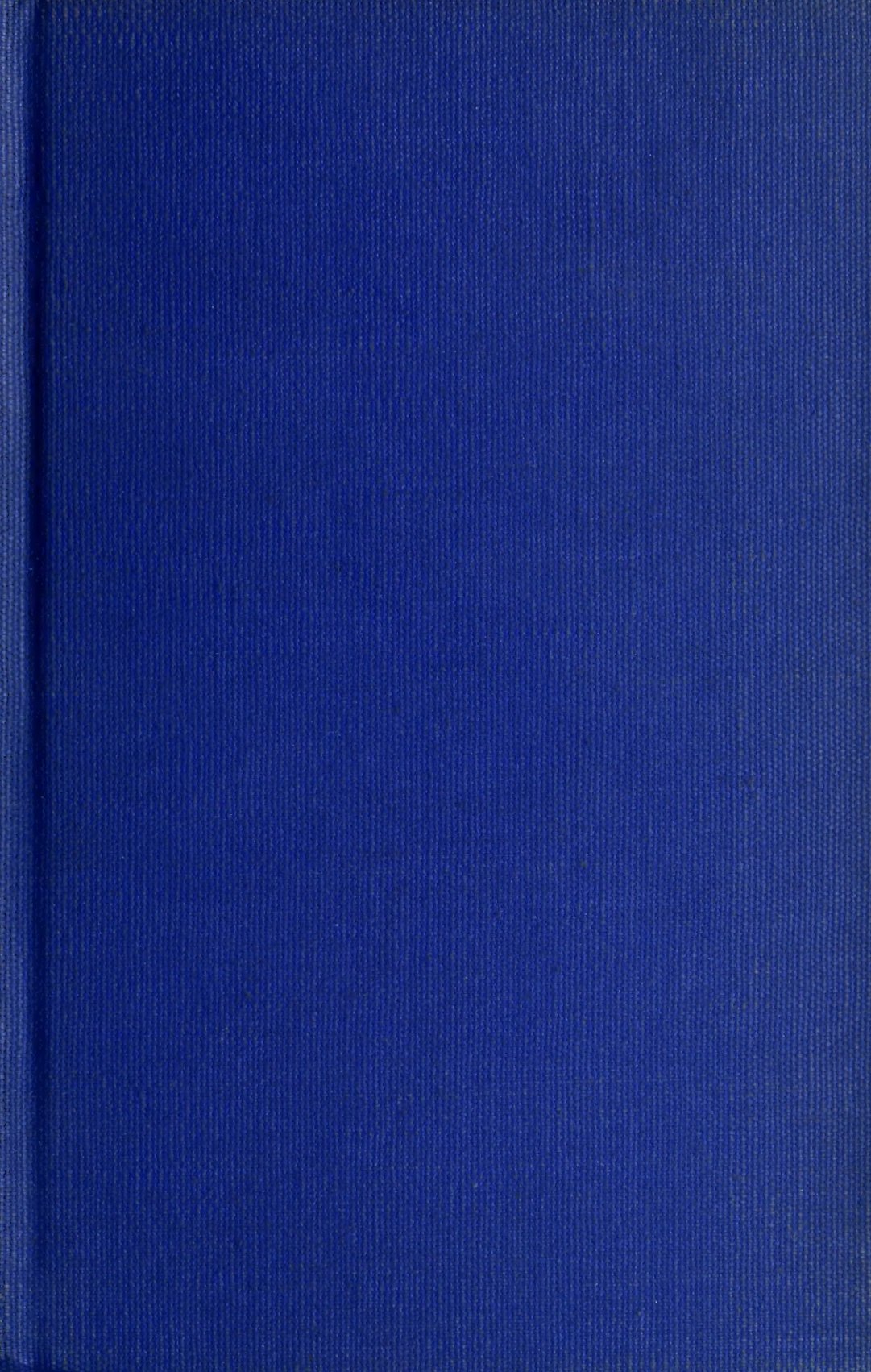
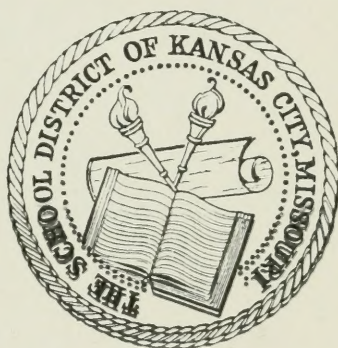




Kansas City
Public Library



This Volume is for
REFERENCE USE ONLY

YRABBU LUBU
YTO BABAY
OH

PUBLIC LIBRARY
KANSAS CITY
MO

THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

ADVISORY BOARD

S. BRACKEN

M. J. KELLY

F. R. KAPPEL

E. J. McNEELY

EDITORIAL COMMITTEE

W. H. DOHERTY, *Chairman*

A. J. BUSCH

R. K. HONAMAN

G. D. EDWARDS

F. R. LACK

R. G. ELLIOTT

H. I. ROMNES

J. B. FISK

H. V. SCHMIDT

E. I. GREEN

G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

INDEX

VOLUME XXXIII

1954

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THIS VOLUME IS BOUND IN TWO PARTS

Part I constitutes pages 1-798

Part II constitutes pages 799-1400

KANSAS CITY, MO.
PUBLIC LIBRARY
FEB 1 - 1955

YANKEE CLUB
YTD 22841
ON

LIST OF ISSUES IN VOLUME XXXIII

No. 1	January.....	Pages	1-275
" 2	March.....		277-516
" 3	May.....		517-798
" 4	July.....		799-1022
" 5	September.....		1023-1208
" 6	November.....		1209-1400

Index to Volume XXXIII

A

- AG RELAY *See* Relay(s)
- ABSTRACTS, of Bell System Technical Papers not published in this Journal 260-73, 505-13
- ACCOUNTING *See* Centralized Automatic Message Accounting System
- Almquist, M. L. 1010
- ALPETH CABLE *See* Cable(s)
- ALTERNATE ROUTING
costs 299-301
interoffice trunk networks 279-86
intertoll trunk networks 286-87
principles 278-87
trunk networks, trunk requirements 277-302
- AMPLIFIER(s)
double-stream 1358-60
magnetic
transistors combined with 847
negative resistance and dissected 799-800
"push-pull" DC 849
resistive-wall 1349-50
space-charge wave 1354-55
- ANALOGUE COMPUTER 139-40
- Analysis of Measured Magnetization and Pull Characteristics* (R. L. Peek, Jr.) 79-108
- Anderson, Orson L.
biographical material 1206
Stress Systems in the Solderless Wrapped Connection and Their Permanence 1093-1110
- Anderson, P. W. 1052
- Andrus, J. A. 1052
- Application of Designed Experiments to the Card Translator* (C. R. Brown, M. E. Terry) 369-98
- ARC(s)
initiation 544-45
interrupted 539-42
reversed 543
welding, percussive 888
- Arcing of Electrical Contacts in Telephone Switching Circuits. Part III—Discharge Phenomena on Break of Inductive Circuits* (M. M. Attala) 535-58
- ARMATURE(s)
relay
direction reversion 17
flux build-up 160
force measurement 11
rebounding 17
saturation 53, 95, 99, 104
- Attala, M. M.
Arcing of Electrical Contacts in Telephone Switching Circuits. Part III—Discharge Phenomena on Break of Inductive Circuits 535-58
biographical material 797, 1398
Open-Contact Performance of Twin Contacts, Theory 1373-86
- ATTENUATION COEFFICIENT
(of) circular electric wave in round metallic tubing 1209
- Automatic Contact Welding in Wire Spring Relay Manufacture* (A. L. Quinlan) 897-923
- AUTOMATIC MESSAGE ACCOUNTING
See Centralized automatic message accounting
- AUTOMATIC MOLDING MACHINE 870-72
- AUTOMATIC MULTIPLE HEAD WIRE STRAIGHTENER 863-64

B

- Babbitt, Dr. B. J. 923
- BALLISTIC GALVANOMETER *See* Galvanometer
- BAND FILTER *See* Filter(s)
- Bangert, John T.
biographical material 514

- Transistor as a Network Element* 329-52
- BARIUM TITANATE
crystal
impact forces 12-13
- Beedy, H. 923
- Bennett, W. R. 1010
- BERKELEY TIMER 15
- Blaha, Albert L.
biographical material 1019
Electronic Relay Tester 925-38
- Bozorth, R. M. 1052
- Brannon, Miss M. J. 1194
- Brown, Clayton B.
Application of Designed Experiments to the Card Translator 369-98
biographical material 514
- Brown, D. R. 923
- Brunner, A. J.
biographical material 1019
Wire Straightening and Molding for Wire Spring Relays 859-84
- BUSY HOUR(s)
busy seasons 299
group 294-97
office 294-97
- C
- CAMA *See* Centralized Automatic Message Accounting System
- CABLE(s)
Alpeth 559
coaxial system
nationwide dialing 277
sheath
non-conductive, continuous incremental thickness, measurements of 353-68
polyethylene
manufacture 559-77
telephone cables 353
thickness measurement 559-77
- Stalpeth 559
- CALL(s)
toll
completion 311
- CARD TRANSLATOR *See* Translator
- CARRIER(s)
nationwide dialing 277
- CENTRAL OFFICE (Telephone)
relays 218
- Centralized Automatic Message Accounting System* (G. V. King) 1331-42
- Chase, F. Harold
biographical material 1019
Transistor and Junction Diodes in Telephone Power Plants 827-58
- CHATTER
governor, dial 1292-95
relay, contact, study 17
- CHATTER INTERVAL(s)
relays 17-18
- Cioffi, P. P.
fluxmeter 8
motion of individual domain walls 1052
- CIRCUIT(s)
controller
trunk concentrating equipment 309
elements
coupled lines 661-719
magnetic
multicontact relay, new 1016-17
repeater, E2 1075-79
repeater, E3 1079-80
welding, ideal, design of 892
- CIRCULAR ELECTRIC WAVES *See* Wave(s)
- Clark, G. W. 1052
- Clogston, A. M. 1052
- COAXIAL CABLE SYSTEM *See* Cable(s)
- COIL, TORCIDAL
magnetic field 42
- COMMUNICATION
waveguide as a medium 1209
- COMPUTER(s)
analogue 139-40
- CONFLUENT BAND FILTER *See* Filter(s)
- CONNECTION(S)
solderless wrapped
diffusion stresses 1099-1108
permanence 1093-1110
properties, affected by plating 1108
stress systems 1093-1110
- "CONTACT CHATTER" 17
- Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath* (Wojciechowski) 353-68

- CONTROLLER CIRCUIT** *See* Circuit(s)
CONVERTER(s)
 negative impedance 1055-92
Cosson, H. E.
 biographical material 1019
 Wire Straightening and Molding for Wire Spring Relays 859-84
Coupled Wave Theory and Waveguide Applications (S. E. Miller) 661-719
Cox, Rosemary E.
 biographical material 1398
 electrical contacts, arcing in switching circuits 557
 Open-Contact Performance of Twin Contacts, Theory 1373-86
CROSSBAR SYSTEM(s)
 No. 4A
 card translator 369
 No. 5
 connectors 1111
 relays 2
CRYSTAL (s)
 barium titanate
 impact forces 12-13
 nickel-iron ferrite
 domain walls, motion of 1023-54
CURRENT
 eddy 176-209
 regulators 849
 relays 8, 176-209
- D**
- Davis, Thomas E.**
 biographical material 1019
 Electronic Relay Tester 925-38
Dawson, R. W. 719
DEIONIZATION TIME 547
Dempsey, F. J. 1052
DIAL TELEPHONE, DIALING
 governor, *see* Governor
 nationwide plan 277
 relay, electromechanical 1
 signaling requirements 1309
 toll
 centralized automatic message ac-
 counting system 1331-42
 nationwide 277-302
 operator 277
 switchboard 303
- Diffraction of Plane Radio Waves by a Parabolic Cylinder* (S. O. Rice) 417-504
Dillon, J. F., Jr. 1052
DIODE(s)
 junction
 telephone power plants 827-58
 reference voltage 833-34
 telephone power plants 827-58
 regulator 840
 semiconductor
 negative resistance arising from
 transit time in 799-826
DISCHARGE PHENOMENA 535
DISSIPATION, reduction of 330
DISTORTION, transmission 773, 779
Dixon, J. T. 1010
DOMAIN WALL(s)
 ferromagnetic
 defined 1023
 motion
 (in) nickel-iron ferrite 1023-54
- E**
- Early, James M.**
 biographical material 797
 P-N-I-P and N-P-I-N Junction Transistor Triodes 517-33
EASITRON 1350-52
ECHO(es)
 pulse 752-761
Economics of Telephone Relay Applications (H. N. Wagar) 218-59
EDDY CURRENT
 relays 176-209
ELECTRIC WAVES *See* Wave(s)
ELECTRODE(s)
 welding 885
ELECTROMAGNET(s)
 core 53
 saturation 97-99
 measured magnetization 79-108
 pull characteristics 79-108
ELECTROMAGNETIC FIELD(s)
 shadows cast by hills 439
 toroidal coil 42
 beam
 field equations 400
 magnetically-focusing cylindrical
 wave propagation 399-416

Electronic Relay Tester (A. E. Blaha, T. E. Davis) 925-38
 ELECTROSTATIC GAUGE 925
 ELLWOOD MAGNETOMOTIVE FORCE GAUGE 9
 ENGINEERED ALTERNATE ROUTING 277-78
 Engst, N. K. 884
 Enz, R. E. 1052
 Eppler, Walter T.
 biographical material 797
 incremental sheath thickness measurements 368
 Thickness Measurement and Control in the Manufacture of Polyethylene Cable Sheath 599-77
 ERROR, source and measures 397-98
Error and Control of the Operate Time of Relays: Part I—Theory (R. L. Peek, Jr.) 109-45
Estimation and Control of the Operate Time of Relays: Part II—Design of Optimum Windings (M. A. Logan) 144-86
 EVAPORATION, in welding 885

F

FERRITE(s)
 dielectric constant 1145
 measurements on
 coaxial cable technique 1168
 nickel-iron, motion of individual domain walls in 1023-54
 FERROMAGNETIC DOMAIN WALL(s) *See* Domain Wall(s)
 FIELD(s) *See* Electromagnetic Field(s)
 FILTER(s)
 confluent band 332
 non-inductive 343
 FLUX relays 7-10, 14-15
 FLUXMETER, RECORDING 8-9
 FORCE relay 4-7, 11-12
 Fritsch, W. W. 1330

G

Galt, J. K.
 biographical material 1206
 Motion of Individual Domain Walls in a Nickel-Iron Ferrite 1023-54

GALVANOMETER, BALLISTIC 8
 GAUGE(s)
 electrostatic 925
 Ellwood magnetomotive force 9
 GENERAL TOLL SWITCHING PLAN *See* Switching System(s)
 GLOW-ARC TRANSITIONS 552, 555-56
 GLOW MAINTENANCE 552
 GOVERNOR(s)
 chatter 1292-95
 design, principles 1267-1307
 (for) dial, 7-type 1267-1307
 input torque
 dial speed relation 1290-92
 motion equations 1273-78
 operation 1269-73
 (for) regulating speed of rotary dials 1267-1307
Governor for Telephone Dials—Principles of Design (W. Pferd) 1267-1307
 GRAECO-LATIN SQUARE AND FACTORIAL DESIGNS 369
 Green, E. I. 351
 Guetich, T. H. 1131
Guided-Wave Propagation through Gyromagnetic Media. Part I—The Completely Filled Cylindrical Guide (H. Suhl, L. R. Walker) 579-659
Guided-Wave Propagation through Gyromagnetic Media. Part II—Transverse Magnetization and Non-Reciprocal Helix (H. Suhl, L. R. Walker) 939-86
Guided-Wave Propagation through Gyromagnetic Media. Part III—Perturbation Theory and Miscellaneous Results. (H. Suhl, L. R. Walker) 1133-94
 GYROMAGNETIC MEDIA
 guided-wave propagation through 579-659, 939-86, 1133-94

H

Hamilton, B. H.
 biographical material 1020
 Transistor and Junction Diodes in Telephone Power Plants 827-58
 Hamming, R. W. 1194

HELIX

- cylindrical 969
- non-reciprocal
- transverse magnetization 939-86
- plane 954

Herring, C. 1052

HILLS, SHADOWS BEHIND

- calculation 417-504

Hittinger, W. C. 532

Holden, A. N. 1052

Hopper, H. R. 1052

Hubbard, F. A. 1330

I

IMPEDANCE, negative, theory 1065

IMPEDANCE INVERSION 349

In-Band Single-Frequency Signaling (N.

A. Newell, A. Weaver) 1309-30

INDUCTANCE

- elimination of 342

INSTRON TENSILE TESTING MACHINE

4, 6-7

INTERFERENCE EFFECT(s)

- mode conversion and reconversion 1230-39

INTEROFFICE TRUNK NETWORK(s)

- alternate routing 279-86

Intertoll Trunk Concentrating Equipment

(D. F. Johnston) 303-28

INTERTOLL TRUNK NETWORK(s)

- alternate routing 286-87

ION(s) *See* Deionization Time

J

Jensen, J. F. 1052

Johnson, H. E. 1052

Johnston, Donald F.

- biographical material 514

Intertoll Trunk Concentrating Equipment 303-28

JUNCTION DIODE *See* Diode(s)

JUNCTION RECTIFIER(s) *See* Rectifier(s)

JUNCTION TRANSITOR *See* Transistor(s)

K

Kahl, H. 1092

Keller, Arthur C.

- biographical material 274

Design of Relays 1-2

Kessel, R. L. 923

King, Gerald V.

- biographical material 1398

Centralized Automatic Message Accounting System 1331-42

KLYSTRON 1348-49

Koehler, D. C. 923

Kompfner, R. 1194

L

Lambert, Mrs. C. A.

- guided wave propagation through gyromagnetic media 1194

wave propagation along an electron beam 416

LANDAU-LIFSHITZ EQUATION 1023

Lewis, J. A.

- biographical material 514

Wave Propagation along a Magnetically-Focused Cylindrical Electron Beam 399-416

Llewellyn, F. B.

- pulse transmission, theoretical fundamentals 1010

repeaters, negative impedance 1092

Logan, Mason A.

- biographical material 274

Estimation and Control of the Operate Time of Relays: Part II—Design of Optimum Windings 144-86

slow release relay design 217

LONG DISTANCE WAVEGUIDES *See*

Waveguide(s)

Lucek, C. W. 1330

Lundry, W. R. 351

M

McConnell, J. H. 938

McGlasson, J. 532

Madden, J. J.

- contact welding in wire spring relays 923

percussive welding 895

MAGNET(s) *See* Electromagnet(s)

Magnetic Design of Relays (R. L. Peek, Jr., H. N. Wagar) 23-78

MAGNETIC FIELD(s) *See* Electromagnetic Field(s)

- MAGNETIC MATERIAL(s)
 properties 37-39
 Mandeville, D. 719
 Mason, W. P.
 biographical material 1206
 *Stress Systems in the Solderless
 Wrapped Connection and Their
 Permanence* 1093-1110
 Matthias, B. T. 1052
 MAXWELL'S EQUATION 1168
 Mazza, L. L. 884
 Meola, L. P. 532
 Merrill, J. L., Jr.
 biographical material 1206
 *Negative Impedance Telephone Re-
 peaters* 1055-92
 Meyers, S. T. 1092
 MICROWAVE(s)
 wavelength 417
 MICROWAVE TUBE(s)
 operation
 in terms of waves 1343
 using long electron beam
 low-level or small-signal behavior
 1343-72
 wave picture 1343-72
 Miller, Stewart E.
 biographical material 798, 1399
 *Coupled Wave Theory and Waveguide
 Applications* 661-719
 *Waveguide as a Communication Me-
 dium* 1209-65
 MILLIMETER WAVE(s)
 techniques 1253-56
 Millman, S. 1052
 Mitchell, G. A. 923
 Mitchell, Miss L. 895
 MOLDING
 automatic wire spring relay block as-
 semblies 870-83
 die 874-77, 879
 wire 879-81
 Moore, C. R.
 dial, speed of 1266
 Moore, H. R. 1052
 Morgan, S. P., Jr.
 guided wave propagation through
 gyromagnetic media 1194
 percussive welding 895
 Morton, J. A. 532
*Motion of Individual Domain Walls in a
 Nickel-Iron Ferrite* (J. K. Galt)
 1023-54
*Multicontact Relay, New, for Telephone
 Switching Systems* (I. S. Rafuse)
 1111-32
 Mumford, W. W. 719
- N
- NATIONWIDE DIALING *See* DIAL TELE-
 PHONE
 Neel, R. I. 368
 NEGATIVE IMPEDANCE *See* Impedance
 NEGATIVE IMPEDANCE CONVERTER *See*
 Converter(s)
Negative Impedance Telephone Repeaters
 (J. L. Merrill, Jr., A. F. Rose,
 J. O. Smethurst) 1055-92
 NEGATIVE RESISTANCE *See* Resistance
*Negative Resistance Arising from Transit
 Time in Semiconductor Diodes* (W.
 Shockley) 799-826
 NETWORK(s)
 interoffice, trunk
 alternate routing 279-86
 intertoll trunk
 alternate routing 286-87
 transistors 329
 trunk
 alternate routing, traffic engineering
 techniques for determining trunk
 requirements 277-302
 Newell, Norman A.
 biographical material 1399-1400
 In-Band Single-Frequency Signaling
 1309-30
 NICKEL-IRON FERRITE *See* Ferrite(s)
 NOISE
 cancellation 1361-62
 deamplification 1360-61
 N-P-I-N JUNCTION TRANSISTOR *See*
 Transistor(s)
- O
- OFFICE busy-hour 294-97
*Open-Contact Performance of Twin Con-
 tacts, Theory* (M. M. Atalla, Miss
 R. E. Cox) 1373-86

OPERATE TIME

(of) relays 109-86

OPERATOR TOLL DIALING 277

OSCILLOSCOPE, cathode ray 17

O'Toole, J. L. 369

P

PARABOLIC CYLINDER

plane radio waves, diffraction of
417-504

Paulson, C. 884

Peek, Robert L., Jr.

*Analysis of Measured Magnetization
and Pull Characteristics* 79-108

biographical material 275

economics of telephone relay applica-
tions 256

electronic relay tester 938

*Estimation and Control of the Operate
Time of Relays: Part I—Theory*
109-43*Magnetic Design of Relays* 23-78*Principles of Slow Release Relay
Design* 187-217PERCUSSIVE WELDING *See* Welding

PERTURBATION METHODS 1134-68

Peterson, J. W. 532

Pferd, William

biographical material 1398-99

*Governor for Telephone Dials—Princi-
ples of Design* 1267-1307PHENOLIC RESIN *See* Resin

Pierce, John R.

biographical material 1399

Wave Picture of Microwave Tubes
1343-72wave propagation along an electron
beam 416*P-N-I-P and N-P-I-N Junction Tran-
sistor Triodes* (J. M. Early) 517-33*P-N-I-P JUNCTION TRANSISTOR See*
Transistor(s)

POLDER RELATION 1172

POLYETHYLENE

cable sheath thickness 559

POSITIVE RESISTANCE *See* Resistance

POWER

*P-N-I-P and N-P-I-N junction tran-
sistor triodes* 517

POWER PLANTS (Telephone)

junction diodes 827-58

transistors 827-48

Posin, M. 327

Principles of Slow Release Relay Design
(R. L. Peek, Jr.) 187-217*Problems in Percussive Welding* (E. E.
Sumner) 885-95PULSE TRANSMISSION *See* Transmission

Q

Quate, C. F.

wave propagation along an electron
beam 416

Quinlan, A. L.

*Automatic Contact Welding in Wire
Spring Relay Manufacture* 897-923
biographical material 1020

R

RADIO

propagation over a succession of
ridges on the earth's surface 438
short waves

hills, the effect of 417-504

Rafuse, Irad S.

biological material 1207

*New Multicontact Relay for Telephone
Switching Systems* 1111-32

Randall, Mrs. K. R.

economics of telephone relay applica-
tions 256

slow release relay design 217

Read, W. T., Jr. 1052

Rebarber, Mrs. A. 1194

RECORDER(s)

X-Y 5, 8-9

RECORDING FLUXMETER *See* Fluxmeter

RECTIFIER(s)

germanium, 65-volt, 200-ampere 855
junction 828-31

magnetic

regulated 854

thyatron tube 851-54

REFERENCE VOLTAGE DIODE *See* Di-
ode(s)

REGULATOR(s)

current 849

series 844

- series current 846
- shunt 840
- shunt transistor 842
- simple series 845
- voltage 848
- RELAY(s)
 - AG 187-217
 - all busy 321-22
 - armature
 - direction reversion 17
 - flux build-up 160
 - force measurement 11
 - rebounding 17
 - saturation 53, 95, 99, 104
 - central office, telephone 218
 - chatter intervals 17-18
 - coil characteristics 25-27
 - current 8
 - demagnetization relations 35-37
 - design of 1-259
 - dynamic measurements 10
 - economics 218-59
 - eddy currents 176, 209
 - electromechanical 1
 - end 314-15
 - end of cycle 321-22
 - failure 1
 - flux 7-10, 14-15
 - gauging a 935-37
 - general purpose 14
 - group restore 321-22
 - high speed 11, 53
 - high speed wire spring multicontact 1111
 - identical 169
 - magnet 4-9, 23
 - magnetic design 23-78
 - manufacturing costs 218-59
 - measuring equipment 3-22
 - multicontact, new
 - (for) switching systems 1111-32
 - operate time, estimation and control theory 109-43
 - windings, optimum, design 144-86
 - release time 178-84, 187-217
 - release waiting time 133-38
 - series 168
 - series connected 170-76
 - single 146
 - slide between parts 12
 - slow release
 - principles of 187-217
 - "tens gate" 303
 - "tens group" 313
 - tester
 - electronic 925-38
 - time 15-21, 154-60, 164-66
 - trunk advance 320
 - two like parallel relays in series with a third relay 170
 - unit gates 314-15
 - winding(s) 144-86
 - cost 223
 - wire spring
 - block assemblies, automatic molding 870
 - contact welding 897-923
 - nickel silver wire 859
 - phenolic resin blocks, molding 870-84
 - silicon copper wire 859
 - springs 897
 - welding, automatic contact 897-923
 - wire straightening and molding 859-84
- Relay Measuring Equipment* (H. N. Wagar) 3-22
- Relays, Design of* (A. C. Keller) 1-2
- RELEASE TIME
 - relays 178-84
- REPEATER(s)
 - negative impedance 1055-92
 - E1 1055
 - E2 1055-92
 - E3 1055-92
 - series-type, application 1056-58
- RESIN (phenolic)
 - molding, for wire spring relay blocks 870-84
- RESISTANCE
 - negative 330-32
 - positive 330
- Rice, Stephen O.
 - biographical material 515
 - Diffraction of Plane Radio Waves by a Parabolic Cylinder* (Stephen O. Rice) 417-504

- Rigrod, William W.
 biographical material 515
Wave Propagation along a Magnetically-Focused Cylindrical Electron Beam 399-416
- Roach, J. M. 923
- Robb, D. T. 368
- Rose, Arthur F.
 biographical material 1207
Negative Impedance Telephone Repeaters 1055-92
- ROTTING, OF TELEPHONE TRAFFIC, ALTERNATE 278-87
- trunk networks, trunk requirements 277-302
- Rowen, J. H. 1194
- Ryder, R. M. 532

S

- SF *See* Single-Frequency
- Schelkunoff, S. A.
 circular electric wave, loss coefficient 1210
- Schenck, A. K. 1330
- Schultz, F. A. 884
- Seider, J. P. 923
- SEMICONDUCTOR(s), SEMICONDUCTING MATERIALS
- diodes
 negative resistance arising from transit time 799-826
 junction transistor triodes 517
- SHADOWS, BEHIND HILLS
 calculation 417-504
- Shafer, W. L., Jr. 326
- SHEATH *See* Cable: sheath
- Shockley, William
 biographical material 1020
Negative Resistance Arising from Transit Time in Semiconductor Diodes 799-826
- SHUNT REGULATOR(s) *See* Regulator(s)
- SHUNT TRANSISTOR REGULATOR(s) *See* Regulator(s)
- SIGNAL(s), SIGNALING SYSTEMS
 (for) intertoll telephone trunks 1309-12
 single-frequency
 in-band 1309-30
- SIGNAL-LOSS EFFECT
 mode conversion and reconversion 1229
- SINGLE-FREQUENCY SIGNALING *See* Signal(s)
- Smethurst, J. O.
 biographical material 1207
Negative Impedance Telephone Repeaters 1055-92
- Smith, Donald H.
 biographical material 1021
Transistor and Junction Diodes in Telephone Power Plants 827-58
- Soffel, R. O. 1330
- SOLDERLESS WRAPPED CONNECTIONS
 See Connection(s)
- Southworth, G. C.
 circular electric wave, loss coefficient 1210
- Spillar, R. 923
- STALPETH CABLE *See* Cable
- STORAGE
 wire, straightened 861
- STRESS
 hoop, relaxation of 1094-99
Stress Systems in the Solderless Wrapped Connection and Their Permanence (O. L. Anderson, W. P. Mason) 1093-1110
- Strickland, R. W.
 biographical material 1021
Wire Straightening and Molding for Wire Spring Relays 859-84
- Suhl, Harry
 biographical material 798, 1207-08
Guided-Wave Propagation through Gyromagnetic Media. Part I—The Completely filled Cylindrical Guide 579-659
Guided-Wave Propagation through Gyromagnetic Media. Part II—Transverse Magnetization and Non-Reciprocal Helix 939-86
Guided-Wave Propagation through Gyromagnetic Media. Part III—Perturbation Theory and Miscellaneous Results 1133-94
 wave propagation along an electron beam 416

- Sumner, Eric Eden
 biographical material 1021
Problems in Percussive Welding 885-95
- Sunde, Erling D.
 biographical material 798
Theoretical Fundamentals of Pulse Transmission—I 721-88
Theoretical Fundamentals of Pulse Transmission—II 987-1010
- Swickard, A. E. 884
- SWITCHBOARD(s)
 toll
 idle circuits 324
 outward 303
- SWITCHING SYSTEM(s)
 alternate routing methods 287
 circuits
 arcing of electrical contacts 535-58
 common-control type
 special purpose 303
 dial, dependence on 1267
 four-wire 308
 general toll plan 287-90
 intertoll trunk concentrating equipment 303-28
 nationwide dialing 277
 No. 4 type 303
 relay, *see* Relay(s)
 trends 1111
See also Crossbar System(s)

T

- TECHNICAL PAPERS, Bell System, not published in this Journal 789-96, 1011-18, 1195-1205, 1387-98.
- TELEPHONE
 calls, *see* Call(s)
 central office, *see* Central Office
 dial, *see* Dial Telephone
 power plants, *see* Power Plants
 relay, *see* Relay(s)
- TELEPHONE NUMBER(s)
 nationwide plan 277
- TELEPHONE SET
 500-type
 dial, 7-type, governor for 1267-1307

TELEPHONE TRAFFIC

- alternate routing, engineering techniques for determining trunk requirements 277-302
 group busy hour versus office busy hour traffic 294-97
 high usage groups
 location 290
 intertoll trunk concentrating equipment 303-28
 random versus non-random 298
 traffic engineering techniques for determining trunk requirements in alternate routing trunk networks 277-302
See also Busy-Hour(s)
- TEMPERATURE
 junction diodes in telephone power plants 831-32
- TENSILE TESTING 4-6
- Terry, Milton E.
Application of Designed Experiments to the Card Translator 369-98
 biographical material 515
- TEST SET(s)
 (for) relays 925-38
 (for) repeaters 1083-85
Theoretical Fundamentals of Pulse Transmission—I (E. D. Sunde) 721-88
Theoretical Fundamentals of Pulse Transmission—II (E. D. Sunde) 987-1010
Thickness Measurement and Control in the Manufacture of Polyethylene Cable Sheath (W. T. Eppler) 559-77
- THYRATON TUBE RECTIFIER *See* Rectifier(s)
- TIME *See* Operate Time; Transit Time
- Tiner, Mrs. R. M. 1052
- TOLL CALLS *See* Call(s)
- TOLL CROSSBAR SWITCHING OFFICE, first 227
- TOLL DIALING *See* Dial Telephone
- TORCIDAL COIL
 magnetic field 42
- TRAFFIC *See* Telephone Traffic
Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks (C. J. Truitt) 277-302

TRANSISTOR(s)

amplifiers, magnetic, combined with 847

TRANSISTOR(s)

junction

frequency range 517

N-P-I-N triodes 517-33

P-N-I-P triodes 517-33

telephone power plants 827-58

typical 839-40

network element 329-52

shunt regulator, multistage 844

Transistor and Junction Diodes in Telephone Power Plants (F. H. Chase, B. H. Hamilton, D. H. Smith) 827-58

Transistor as a Network Element (J. T. Bangert) 329-52

TRANSIT TIME

semiconductor diodes, negative resistance arising from 799-826

TRANSLATOR

card

designed experiments, application of 369-98

TRANSMISSION

circular electric wave 1211

coupled lines 661-719

distortion 773-779

frequency characteristics 724

measurement, theory of 1085-87

nationwide dialing 277

network

transistors 329

pulse

theoretical fundamentals 721-88, 987-1010

repeaters, negative impedance 1055-92

waveguide 1209

TRIODE(s)

N-P-I-N junction transistor 517-33

P-N-I-P junction transistor 517-33

Truitt, C. J.

biographical material 515

Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks 277-302

TRUNK(s)

intertoll trunk concentrating equipment 303-28

TRUNK NETWORK(s)

alternate routing

trunk requirements 277-302

V

Van Duyne, C. W. 857

VARIANCE

analysis of 395

components of 396

VOLTAGE

P-N-I-P and N-P-I-N junction transistor triodes 517

regulators 848

relays 8, 10

welding 885

W

Wagar, H. N.

biographical material 275

Economics of Telephone Relay Applications 218-59

Magnetic Design of Relays 23-78

Relay Measuring Equipment 3-22

Walker, Laurence R.

biographical material 798, 1208

Guided-Wave Propagation through Gyromagnetic Media. Part I—The Completely Filled Cylindrical Guide 579-659

Guided-Wave Propagation through Gyromagnetic Media. Part II—Transverse Magnetization and Non-Reciprocal Helix 939-86

Guided-Wave Propagation through Gyromagnetic Media. Part III—Perturbation Theory and Miscellaneous Results 1133-94

Walsh, E. G. 1131

WAVE(s)

circular electric

(and) millimeter wave techniques 1253-56

propagation 1210

(in) round metallic tubing, attenuation coefficient 1209

theoretical characteristics 1213-19

- guided
 - gyromagnetic media, propagation through 579-659, 939-86, 1133-94
 - micro-, *see* Microwave(s)
 - millimeter, *see* Millimeter Wave(s)
 - space-charge 1344-48
- Wave Picture of Microwave Tubes* (J. R. Pierce) 1343-72
- Wave Propagation along a Magnetically-Focused Cylindrical Electron Beam* (J. A. Lewis, W. W. Rigrod) 399-416
- WAVEGUIDE(s)
 - (for) circular electric waves, improved forms 1250-52
 - circular, filled
 - field patterns 1175-94
 - communications medium 1209-65
 - copper
 - bending 1209
 - coupled wave theory 661-719
 - dominant-mode 694
 - long distance
 - practicality 1210
- Waveguide as a Communication Medium* (S. E. Miller) 1209-65
- WAVELENGTH
 - shadows that are cast by hills 417
- Weaver, Allan
 - biographical material 1400
- In-Band Single-Frequency Signaling* 1309-30
- Weiss, M. T. 1194
- WELDING
 - automatic contact
 - wire spring relay manufactured 897-923
 - automatic multiple resistance 898
 - contact 897-923
 - evaporation in 885
 - guns 918
 - percussive
 - fundamental problems 885-95
 - (in) wire spring relay manufacture 907-09
 - projection type resistance 899
 - voltage 885
- Wenger, J. A. 532
- WHITTAKER FUNCTIONS 983
- Williams, H. J. 1052
- WIRE(s)
 - anchoring 878
 - bends 859
 - indexing 882
 - molding 879-81
 - shearing 882
 - spooling operation 859
 - straightened
 - storage 861
 - straightener, automatic, multiple head 863-64
 - straightening and molding
 - (for) wire spring relays 859-84
- WIRE SPRING RELAY *See* Relay(s)
- Wire Straightening and Molding for Wire Spring Relays* (A. J. Brunner, H. E. Cosson, R. W. Strickland) 859-84
- Wojciechowski, Bogumil M.
 - biographical material 515
- Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath* 353-68
- polyethylene cable sheath thickness 577
- Worley, O. C. 217
- Wright, J. P. 1052

X

X-Y RECORDER 5, 8-9

THE BELL SYSTEM
7934
7934
Technical Journal

VOTED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII

JANUARY 1954

NUMBER 1

DESIGN OF RELAYS

Introduction	A. C. KELLER	1
Relay Measuring Equipment	H. N. WAGAR	3
Magnetic Design of Relays	R. L. PEEK, JR. AND H. N. WAGAR	23
Analysis of Measured Magnetization and Pull Characteristics	R. L. PEEK, JR.	79
Estimation and Control of the Operate Time of Relays		
Part I—Theory	R. L. PEEK, JR.	109
Part II—Design of Optimum Windings	M. A. LOGAN	144
Principles of Slow Release Relay Design	R. L. PEEK, JR.	187
Economics of Telephone Relay Applications	H. N. WAGAR	218
Symbols		257

Abstracts of Bell System Technical Papers Not Published in this Journal	260
Contributors to this Issue	274

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

- S. BRACKEN, *Chairman of the Board,*
Western Electric Company
- F. R. KAPPEL, *President, Western Electric Company*
- M. J. KELLY, *President, Bell Telephone Laboratories*
- E. J. McNEELY, *Vice President, American Telephone*
and Telegraph Company

EDITORIAL COMMITTEE

- E. I. GREEN, *Chairman*
- | | |
|---------------|---------------|
| A. J. BUSCH | F. R. LACK |
| W. H. DOHERTY | W. H. NUNN |
| G. D. EDWARDS | H. I. ROMNES |
| J. B. FISK | H. V. SCHMIDT |
| R. K. HONAMAN | G. N. THAYER |

EDITORIAL STAFF

- J. D. TEBO, *Editor*
- M. E. STRIEBY, *Managing Editor*
- R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXIII

JANUARY 1954

NUMBER 1

Copyright, 1954, American Telephone and Telegraph Company

Design of Relays

Introduction

(Manuscript received September 28, 1953)

The electromechanical relay is the basic building block of modern dial switching systems and also of various automatic control systems and computers.

All of these systems depend on the action of the relay which is simple in its functions, but has to meet complex requirements placed on it by the systems in which it is used. Perhaps the best illustration of this apparent conflict is to note that a relay has simply to close or open electrical contacts when its coil is energized or deenergized. However, these simple functions must be performed equally well by millions of relays, and each of these must continue to perform reliably for millions and in some cases for more than a billion operations during its lifetime. Furthermore, in many cases the relay must function in a few thousandths of a second, or use little electrical power. The reliability required in telephone switching systems would be considered unreasonable in many other types of equipment and can be judged from the fact that a single failure in 5,000,000 operations is considered to be poor performance.

Stated another way, satisfactory operation for the average relay in modern telephone switching systems is less than one failure in forty years of operation. The measurement of such low trouble rates is, in itself, a difficult and challenging problem.

The need for such a high degree of reliability and for the associated requirement of high speed is evident from a few figures relating to

modern cross-bar switching systems. In these, establishing a single telephone connection between two subscribers makes use, for a very short time interval, of approximately 1,000 relay operations with a total of about 7,000 contacts.

As the telephone plant has become more mechanized with its local and toll dial systems, automatic message accounting, automatic trouble recorders, etc., the relays have been required to do more things with less trouble. Accordingly, the steady progress which has been made in the automatism of switching equipment has depended upon improving the performance of relays. To provide this improvement, relay design must be guided by a clear understanding of the physical relations among all aspects of performance and the construction specified by the design.

Some of the recent work in the area of relay design, production, service and measurements is covered by this issue which is devoted entirely to the analysis and measurement of relay performance, and to the economic considerations which govern optimum relay design. It is evident from the typical statistics given, that the successful and economical operation of switching systems and certain other existing automatic control systems demand the best in relay performance.

A. C. KELLER

Relay Measuring Equipment

By H. N. WAGAR

(Manuscript received September 25, 1953)

The wide variety of technical problems encountered in telephone relay design calls for quantitative measurements of many kinds, involving static performance and dynamic performance. The present article describes some of the more important measuring tools for this purpose, as used in Bell Telephone Laboratories. Instrumentation for evaluation of force, flux, displacement, time, or their combination, is described.

INTRODUCTION

Visitors to the switching development areas in the Bell Telephone Laboratories are often surprised at the extensive amount of measuring equipment used to study so simple a device as the telephone relay. This is because, though the relay itself may be simple, the problems requiring study are extremely complex. Because relays are used in such large quantities, their characteristics affect the economy of the telephone switching system. Not only is their manufacturing cost important; their quality also affects the central office cost. For example, an efficient design lowers the central office power plant cost, and faster-acting relays enable fewer common control units to handle more traffic, which further reduces the cost. These and related objectives pose many complex technical problems involving mechanics, magnetics, kinetics, heating, and the like.

The scope of relay analysis may be judged from the other articles in this issue. In every case there is need for measurement. Sometimes this quantitative work is needed to confirm an analytical relation, sometimes to learn more concerning the basic phenomena, and often to characterize the performance of a test model. This article will describe some of the most-used measuring tools for the study of relays.

Several kinds of measurements are required. In the first place, there are force measurements. Each relay must press the desired number of contacts together with a suitable force. To do this, a magnet must be provided which can move the springs, through whatever distance is required. Thus, measurement is needed of force-displacement characteris-

tics both for contact arrangements, and for magnets. Another field of measurement involves magnetics. Both the mechanical output and the electrical characteristics of the electromagnet are determined by its magnetization relations, requiring the measurement of magnetic flux. Beside such static measurements of mechanical, electrical, and magnetic quantities, corresponding measurements must also be obtained as functions of time in studying relay behavior under the dynamic conditions of actual operation.

The many instruments which furnish such data may be grouped into those for static measurements, and those for dynamic measurements. Some of these tools are described in the following pages, particularly as they relate to force, current, flux, displacement, time, or their combination.

STATIC MEASUREMENTS

The Measurement of Force

For a complete understanding of the operation of relay designs, knowledge of how the forces vary as the magnet air-gap changes, or as the contact members are deflected through their stroke, is of course required. The measurements most needed concern the force versus distance of the contact loads which the magnet must operate; and the force versus air-gap characteristic of the associated magnet. For many years, such measurements were made by a process of hanging weights and setting a micrometer screw, point by point, until complete data were obtained. More recently special instrumentation has made available a pendulum-type tensile tester,¹ and a spring balance device,² which have greatly increased the convenience of making these measurements, point-by-point. Today, however, many such measurements may be made still more conveniently by a machine which automatically causes the gaps to vary and plots a curve of force versus distance. Machines of this type have been developed to a high degree of flexibility for tensile testing of materials such as metals, plastics, textiles, or paper.

One such machine, which is extremely convenient for relay measurements, is the Instron tensile testing machine shown in Fig. 1. This machine is manufactured by the Instron Engineering Corporation of Concord, Mass. In conjunction with suitable current supplies to control the behavior of the relay magnet, and with a few special circuits for automatically correcting for flexure of the relay parts, and making other similar adjustments, this machine has proven eminently suitable for observing all "static" force-deflection characteristics of relays. The manner

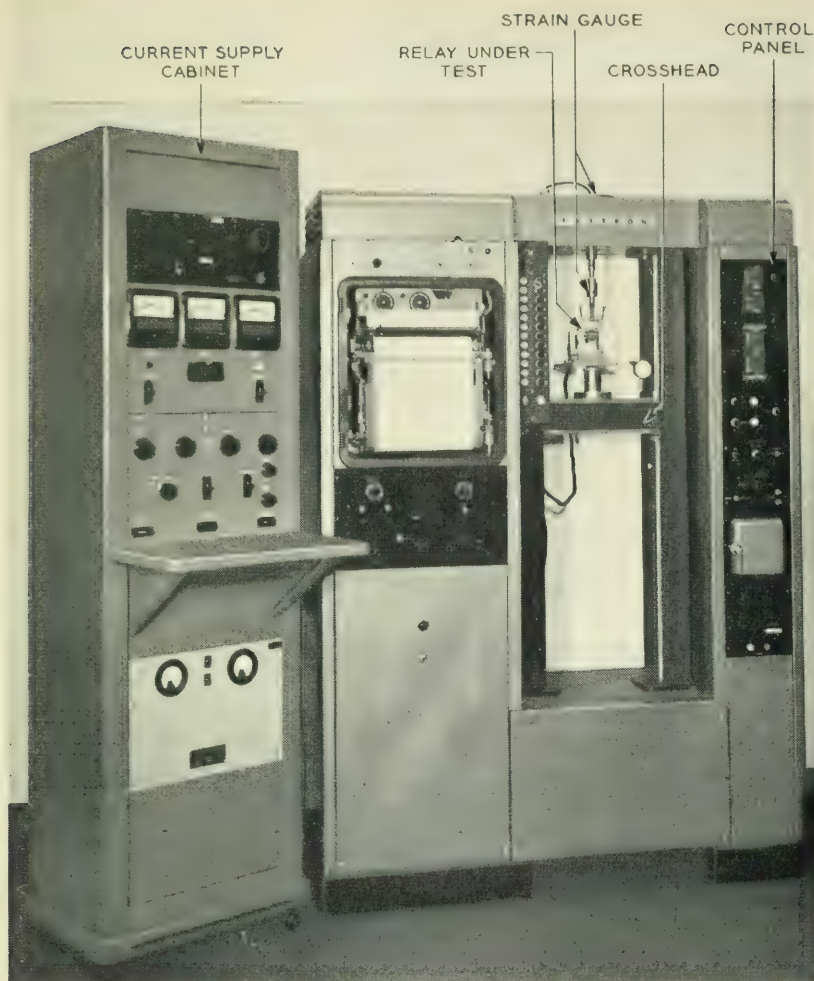


Fig. 1 — Tensile testing machine for force-displacement measurements.

of operation of this machine will be given only in outline as it is described elsewhere.³ The force system to be measured, such as a relay magnet, is mounted on the crosshead of the machine, whose motion up or down may be controlled from the panel on the right. The other side of the force system being measured is connected to a strain gauge fastened to the top framework of the machine, and as the force on the strain gauge progressively changes, an electrical indication is given to the pen of the Leeds and Northrup X-Y recorder which then moves horizontally in proportion

to the force. The crosshead motion is transmitted electrically to a servo system which drives the paper up or down in proportion to the displacement of the mechanism involved. Thus, during the motion of the crosshead, the pen moves proportionately to force, the paper moves proportionately to distance, and a force-deflection curve is drawn. As a result, the tensile characteristics of mechanisms can be recorded very easily. For example, force variations of relay springs or of relay magnets may be measured, or a combination of the two, as desired.

Some typical measurements obtained on this machine are given in Fig. 2 showing on one chart the manner in which electrical contacts change their force against the magnet and the manner in which the magnet pull varies across the gap. Such charts as these, when completely analyzed, enable the designer to choose the proper magnet for a particular spring requirement. Among the useful features of this machine are the accuracy of recording which readily provides one or two per cent accuracy, and the ease of tracing and retracing a particular measurement. Also readily obtained are exact information on the mechanical hysteresis losses within the mechanism. For example, in curve 1 of Fig. 2, the load characteristic of the contact springs is seen to have two values. The upper one represents the force on the closure stroke while the lower one represents the force on separation. The area between these two curves is the hysteresis loss; friction at any point is readily estimated as one-half the difference between upper and lower force readings. Special information, such as the exact location of first closure or first separation of contacts can also be included, as the machine is provided with a circuit to insert a "pip" on the record when desired.

This instrument has marked advantages because of its accuracy, speed, convenience and wide range of uses. With almost equal ease, mechanisms whose full force variation is 2 grams, and those ranging up to half a ton can be measured. Once a suitable jig has been prepared to properly mount the parts, the Instron machine will furnish a complete set of data in less than an hour, whereas by previous methods it would have required a matter of days.

The time required for each individual curve in the figure is the time allowed for the crosshead to move through the relay stroke. This usually is about 30 to 60 seconds, which adequately simulates the static characteristics of the relay force system; i.e., those force properties that would be measured at a particular point if the armature were held there. These static characteristics are commonly used in relay design to establish performance criteria for ordinary relays. However, in detailed studies of relay dynamics, these static characteristics must be supplemented by

estimates or measurements of the inertia and other forces which also occur. The direct measurement of dynamic performance is described in a later section.

The Measurement of Flux

The mechanical output of a magnet depends on its magnetic quality. As is shown in companion articles in this issue, it is important for the designer to know the values of the closed gap reluctance $\mathcal{R}_0$, the leakage reluctance $\mathcal{R}_L$, and the effective pole face area A . Beside these important

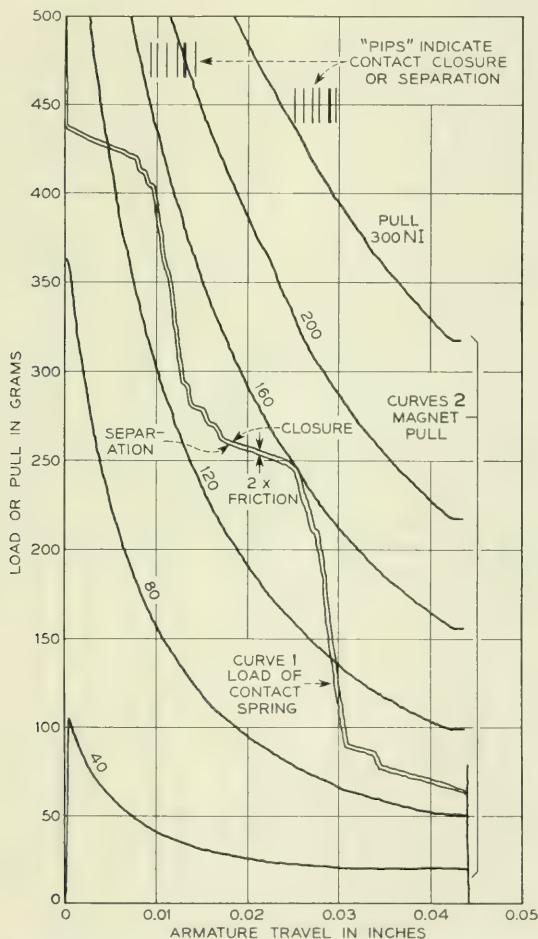


Fig. 2 - Typical recording of relay performance made on Instron machine.

figures of merit, values for saturation flux, ϕ'' , and leakage flux ratios at various points in the magnetic circuit, are commonly needed. For such information, flux as a function of ampere turns must be measured. Perhaps the most effective method is to make determinations of the average flux in the magnet structure at a number of different air gaps; then by analysis determine values for the constants just enumerated. A series of magnetization curves for the relay magnet are required similar to those commonly taken on magnetic "ring samples," except that curves are obtained for several different air gaps. The ballistic galvanometer is the most familiar instrument for this purpose, still used for certain problems. In the Bell Laboratories, however, an extremely versatile recording fluxmeter was developed some ten years ago by P. P. Cioffi for use in research on magnetic materials.⁴ One of these instruments is now in constant use for relay measurements because of its accuracy, versatility and the rapidity with which tests can be made. The complete unit is shown in Fig. 3, being made up of a power supply system, galvanometer-integrator unit, and X-Y recorder. Its operation will now be described very briefly.

The magnet to be tested is provided with a search coil. This search coil may be in the form of a winding physically in parallel with the main supply winding, a coil of few turns uniformly spread over the outside of the coil, or of a specially wound coil mounted at a particular point of interest. When current is progressively varied through the main winding, the voltage induced in the search coil by the magnet flux is transmitted to the galvanometer whose associated optical system divides a light beam between two photocells. The resulting photocell voltage unbalance is amplified and coupled back into the search coil circuit through a mutual inductance, so poled as to tend to restore the galvanometer deflection to zero. The feedback current in the mutual circuit is proportional to the flux, and is used to drive the pen on the X-Y recorder. The paper is driven proportionately to the current in the winding, with the result that flux versus current (or ampere turns) is plotted directly. The instrument gives accuracy of $\pm 1_2$ per cent over a wide range of fluxes and currents, with provisions made for readily changing the scale to cover different windings, operating voltages or magnet shapes. Measurements are made in a few minutes compared to a considerably greater effort when using the ballistic galvanometer. Auxiliary features are also provided permitting convenient and complete demagnetization of the test magnets between readings. Readings are extremely stable and repeatable. Typical results are shown in Fig. 4 where a series of magnetization curves for a particular test magnet are shown. By



Fig. 3 — Recording fluxmeter.

methods of cross-analysis described in a separate article in this issue,⁵ these data may be used to find the figures of merit for the magnet, and other information concerning its performance.

Often the measurement of magnetic potential at different locations in a magnetic circuit is required. Among the more versatile of such tools is the Ellwood magnetomotive force gauge.⁶ A brief description of its operation appears in a companion article.⁶

The above measurements provide information on the static magnetic characteristics of a relay, but they do not provide all the information

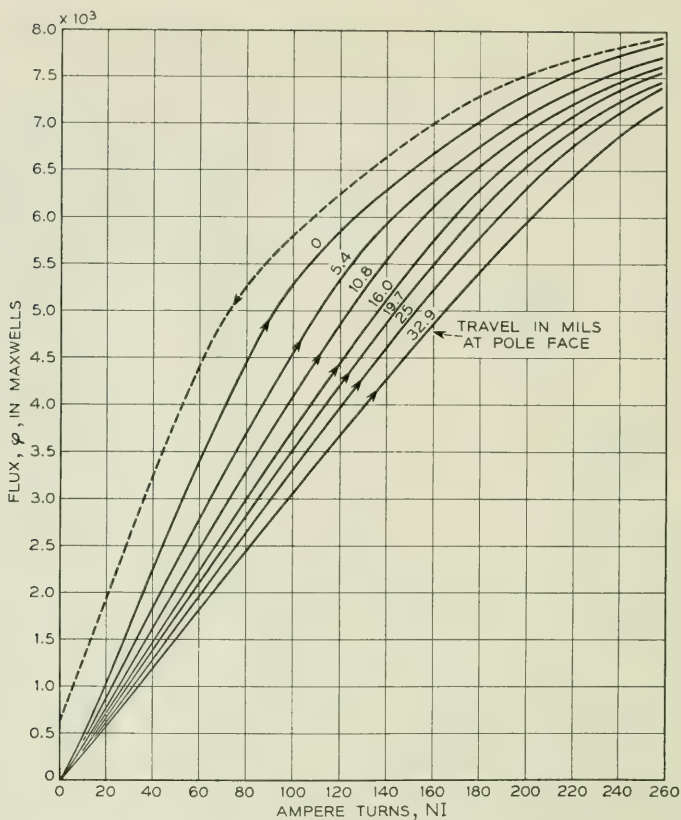


Fig. 4 — Typical recording of flux-ampere turn data on a relay magnet.

needed to describe the action while parts are in motion. For measurements involving dynamics, additional methods are described below.

DYNAMIC MEASUREMENTS

The measurements just described are indispensable in obtaining basic information on the relay's ability to perform its prescribed mechanical functions, which in turn depend upon its magnetic characteristics. They are needed in the every-day calculation of winding designs and the application of contact spring combinations which will be reliably operated by a given magnet. Cases arise in service, however, where very rigid requirements must be placed on closure and separation to insure the desired performance. Complete knowledge of the influence of voltage variations, mechanical adjustments, numbers of springs and their natural fre-

quencies is needed. Often impacts and slide of parts must be minimized in order to reduce mechanical wear. Furthermore, the design of high speed relays requires an understanding of how the flux rises and decays when current is connected or disconnected.

For the experimenter in these fields, many tools are available covering measurements of force, flux, current, and time, or combinations thereof. These measurements may involve shadowgraph, high speed motion picture, or various transducer techniques, and many special methods; some of these are described below.

The Measurement of Force

There is a definite need to know how forces in a relay structure vary with time. It would, for example, be very useful to measure the force experienced by the relay armature as it presses against its spring load during operation. While a satisfactory measurement technique for this problem has not yet been found, there are a number of similar type measurements which can be applied to individual mechanical problems in the structure. For example, with the higher-speed relay functions of today, it is found that relays made by older methods suffer excessive

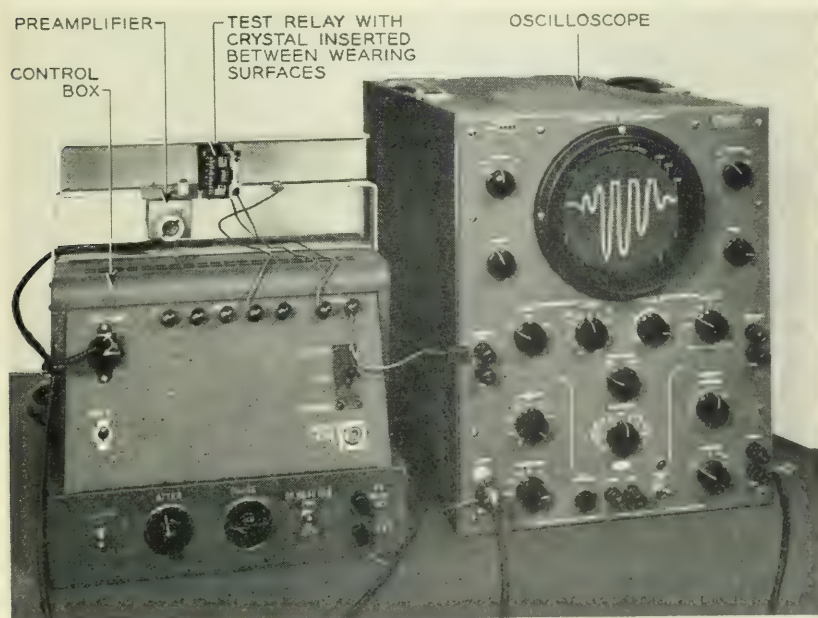


Fig. 5 — Barium titanate test set.

mechanical wear. In order to decide whether this is due to impacts, slide of parts, or vibration, methods have been developed to isolate some of the transient forces encountered in relay operation, and measure them individually.

Rapidly varying forces, such as impact forces, may be measured with the barium titanate crystal, which acts piezoelectrically to yield a voltage across its surfaces when subjected to compression or shear type forces. Since it can be obtained in very small sizes, it may be mounted within the relay as a substitute for the part to be studied. By observing the voltages that have been developed across it during relay operation, one may readily measure the forces involved.⁷ Such an arrangement can be made to provide a faithful frequency response over a wide range, though it calls for careful amplifier circuit design. The unit shown in Fig. 5 provides for accurate frequency response for impact forces varying as high as 50 kilocycles. Fig. 6 shows a typical force-time relation obtained with this equipment. In some recent measurements on an experimental relay, the impact forces between the driving members were measured to be approximately five-fold the static force, but not sufficient to explain pulverizing and other damage to certain of the functioning parts.

Imposed Motion

Slide between parts has been found to be a very damaging source of wear in telephone relays. It has been studied in some detail by means of

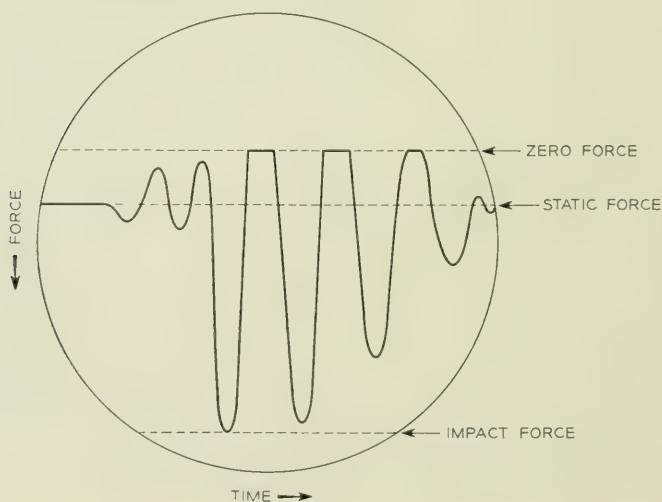


Fig. 6 — Oscillogram of impact in a relay.

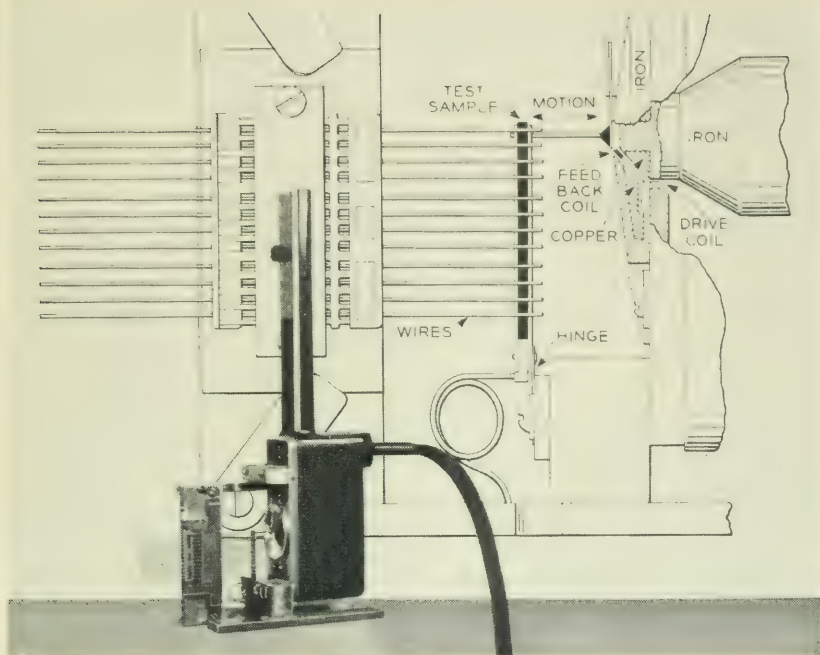


Fig. 7 -- Photo of 1A recorder setup.

transducer elements whose motion is controllable to simulate a wide range of slide conditions typical of possible designs, from which wear measurements may be taken and analyzed. Such methods were also described in the previous reference.⁷ The measurements may be made either by means of a barium titanate element or by means of a phonograph recording head such as the Western Electric 1A recorder. Because of the feedback coil and amplifier circuit, the latter transducer will generate the precise amplitude of motion desired for any particular test merely by properly setting its input current. It is normally driven at frequencies in the order of 1,000 cycles, permitting a large number of slides to be obtained in a short time and affording information to the designer on a very accelerated basis. Through the choice of different amplitudes, different materials to be tested, and different forces between them, a wide range of variables can be covered, to guide the designer in the proper choice of materials and conditions of use. A recorder setup for a typical test is shown in Fig. 7; data taken on a typical set of parts for an AF type relay are shown in Figure 8.

The Measurement of Flux

When attempts were first made to extend the speed performance of general purpose relays much below 8 or 10 milliseconds functioning time, it was found that the simplified eddy-current theory, which treated the core like a single short-circuited turn, was no longer sufficiently accurate. The subject has been clarified through the use of the dynamic fluxmeter which permits a direct determination of how the relay flux changes with time, under actual conditions. This fluxmeter, originally due to E. L. Norton and recently refined by M. A. Logan, was described in a previous issue,⁸ and has proven most useful for determining those constants of the relay which the designer must understand in order to work much below 8 milliseconds functioning time. The instrument is shown in Fig. 9. It requires that the relay under test be pulsed, usually at a speed of about 10 to 20 cycles. When the main winding is alternately connected and disconnected, the search coil which surrounds the winding sends alternate positive and negative voltage impulses to a dc ammeter connected across its terminals. Now it can be shown that when the search coil and meter are disconnected during the interval from time zero to the moment of interest, then the resulting reading on the meter is exactly proportional to the flux at that moment. This switching function is provided by a timing circuit comprising a 40-ke frequency source driving a 3 decade counting ring, under control of switches permitting selection of the number of cycles of delay time. In this way the flux may be measured at intervals of 25 millionths of a second, with better than 1 per cent accuracy.

One interesting measurement has shown that in the short time intervals of present-day relays, eddy currents have less effect on the initial

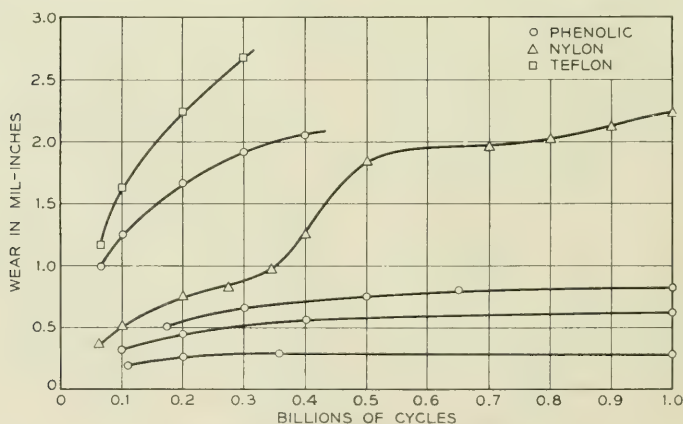


Fig. 8 — Slide data on AF relay.



Fig. 9 — Photo of dynamic fluxmeter.

development and decay of the field than was previously thought to be the case.

The Measurement of Time — Electrical Effects

The actual functioning time of the relay is one of the more important behavior characteristics which must be tabulated for the particular design. Its measurement can be quite simple or quite complex depending on the problem involved.

One of the simpler time measurements is that permitted by means of the arrangement shown in Fig. 10. This particular equipment is composed of a control circuit for studying the various relay timing conditions, and a Berkeley Timer made by the Berkeley Instrument Company. The timer is provided with electronic controls which cause it to start counting cycles from a standard frequency when given an electrical impulse and to stop counting upon receiving a second impulse. The circuit may be arranged to give the first impulse when the relay winding is connected or disconnected and the second impulse when the contacts or other

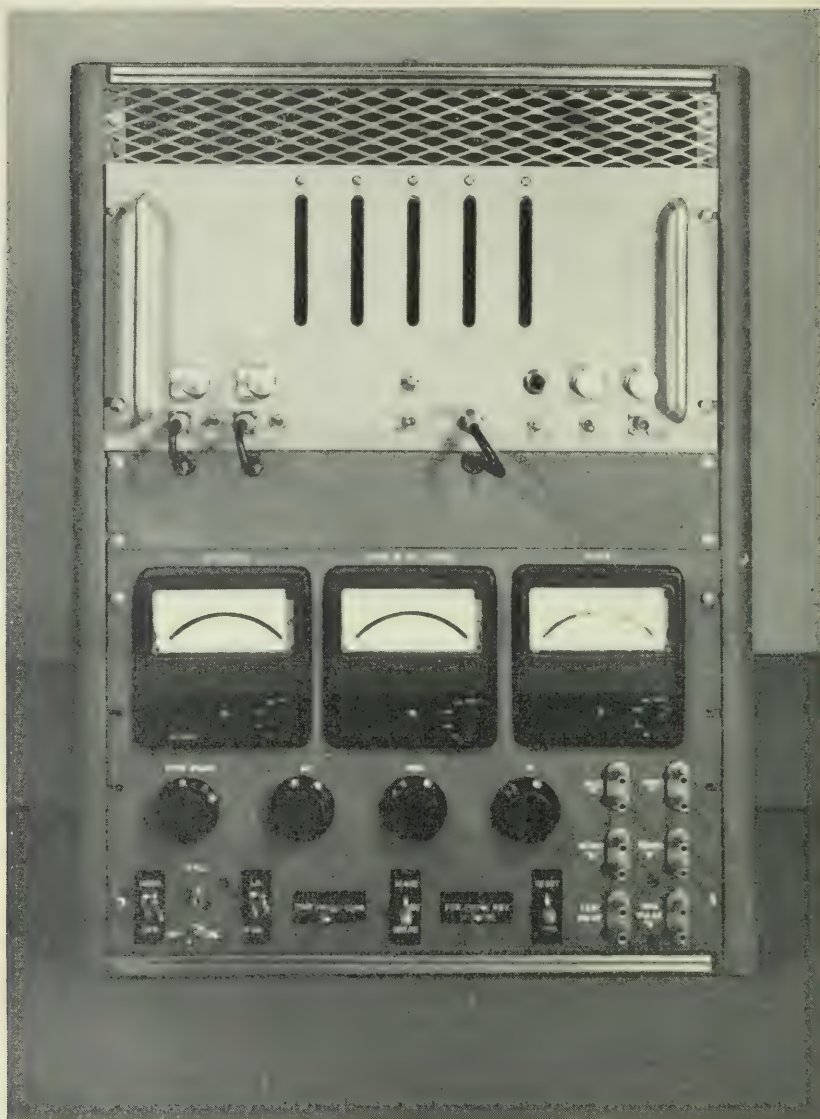


Fig. 10 — Photo of timing setup.

members either touch or separate. It is timed by a 100-ke oscillator allowing measurements to be made at intervals of 10 millionths of a second. It is in constant use for the every-day measurement of relay timing.

The simple timing measurements just described afford data on the

basic timing characteristics of the structure, but may give insufficient information on the timing performance of the relay contacts in a particular circuit. For example, contacts may close as desired but then reopen intermittently in a manner to cause either unwanted circuit behavior or undesirable contact sparking and consequent erosion. Sometimes the relay armature, in crossing the air-gap, encounters sudden loads as springs are successively picked up in the stroke, causing it to momentarily reverse its direction before pulling home. Also, when the relay armature releases, it may strike the backstop and rebound, recrossing a portion of its gap and causing undesirable reactuating of the contacts. Such momentary opening and closing of the contacts is classed as "contact chatter," and special means are needed to detect, measure, and understand its behavior.

The string oscillograph has been extremely useful for recording and studying such timing effects when the contact chatter was of comparatively long duration and low frequency. For relays whose functioning time exceeded 10 milliseconds and where chatter intervals were in the order of 2 or 3 milliseconds each for possibly 6 or 7 successive times, this instrument was most effective. For many of the faster relays, however, where chatter of this type has been completely eliminated, there remain problems of much higher frequency, shorter duration chatter, which are still of great importance to the circuit designer. In such cases the cathode ray oscilloscope is used.

A cathode ray oscilloscope arrangement for the study of chatter is shown in Fig. 11. The horizontal axis of the oscilloscope has a calibrated time base and the trace is displaced vertically to mark closing of the contacts or other events of interest. The horizontal sweep may be triggered at the initial closure of the contacts, but a variable time delay

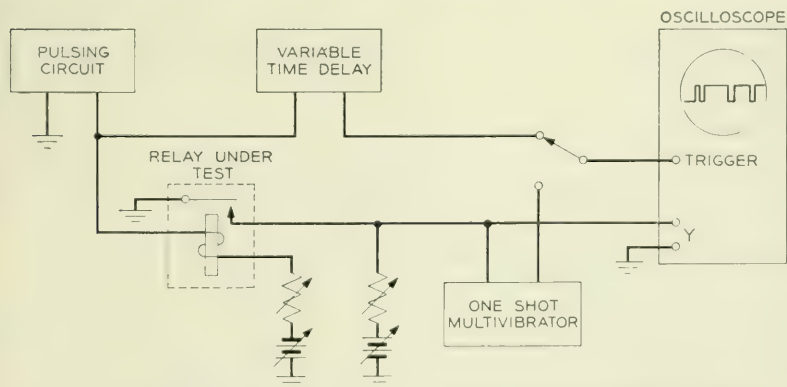


Fig. 11 — Cathode ray oscilloscope circuit for the study of contact chatter.

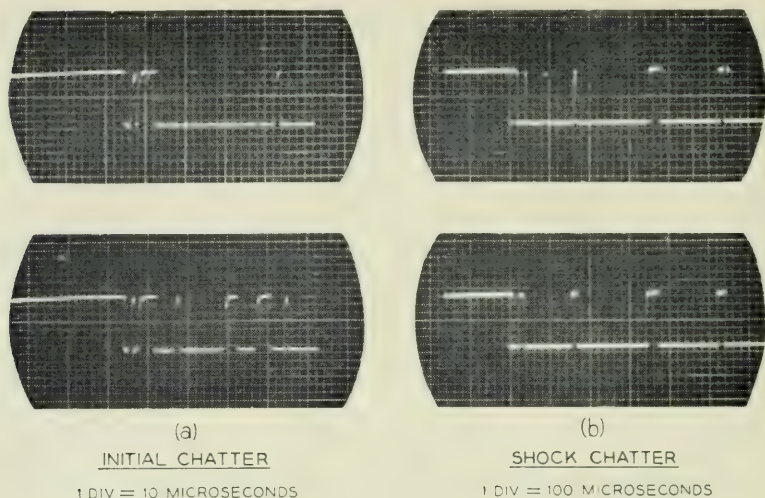


Fig. 12 — Views of chatter obtained by oscilloscope.

can be used to trigger the sweep at a later time so that any portion of the chatter can be observed in detail. Typical measurements are shown in Fig. 12. In each of the views, the horizontal lines starting at the upper left represent open contacts, the lower horizontal lines represent closed contacts, and any jumping between them indicates chatter. Fig. 12(a) shows the relatively high frequency chatter which occurs immediately after contacts close. This is called initial chatter and is caused by vibration of contact springs in their higher modes due to the impact of mating contacts. Initial chatter differs in character from the lower frequency shock chatter shown in Fig. 12(b). Shock chatter is caused by wire vibrations induced by the shock of the armature striking the core. The time of each reopen of the contacts and the time between opens is much longer than for initial chatter.

Many measurements of the type just described are made in the course of a relay development. They enable the relay designer to relate chatter performance to design characteristics of relays so that contact chatter can be reduced or eliminated. Such measurements also allow one to classify chatter performance according to the circuit application.

The Measurement of Time — Mechanical Effects

Electrical timing measurements previously described give data on over-all effective performance; to understand these results one often needs to measure displacement-time characteristics both alone, and in relation to current versus time and flux versus time. The string oscillo-

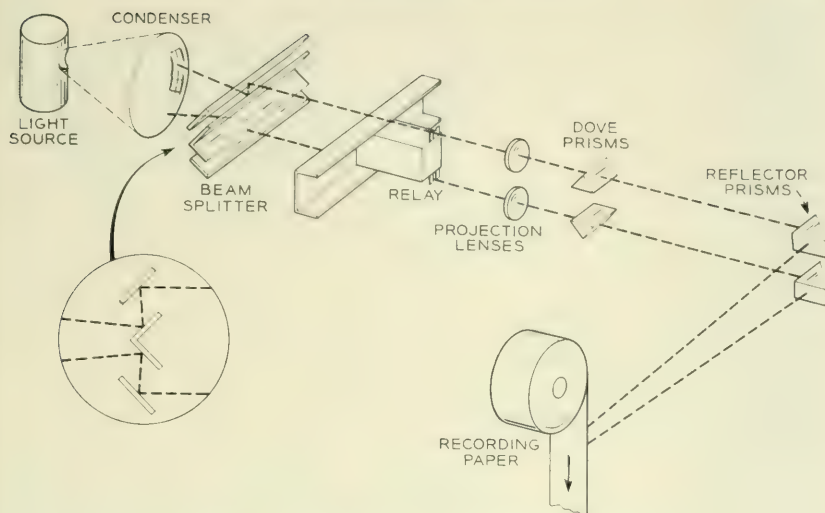


Fig. 13 — Schematic of shadowgraph action

graph mentioned above, has on many occasions been modified to also record a silhouette of the moving parts of the structure on a moving strip of photographic paper, and thus provide a simultaneous trace of displacement and current against time. One of the more advanced shadowgraph-oscillograph equipments is shown in Fig. 13, and a typical record made on a wire spring relay is given in Fig. 14. In this case the single record shows the current to the winding and the mechanical movement at the two opposite ends of the contact-operating card attached to the armature of this relay. From such measurements, velocities of impact, location of the parts at a given instant, stagger between parts, relative motion, unbalanced motion, and the like, may all be determined and correlated with electrical changes.

Motion of parts may also be studied optically to give displacement or velocity data, using the photocell. Properly placed flags on the moving

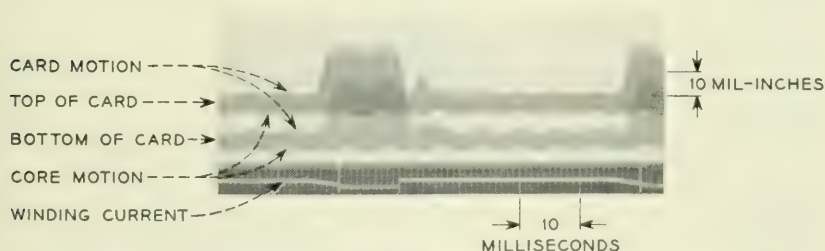


Fig. 14 — Typical shadowgram of 287 relay armature motion.

TIME IN MILLISECONDS

0

1

2

3

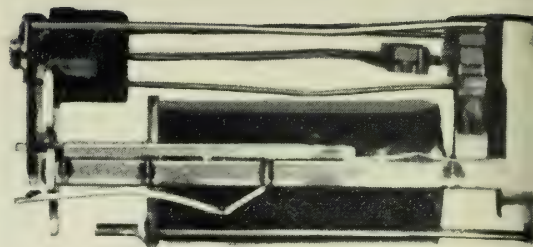
4

5

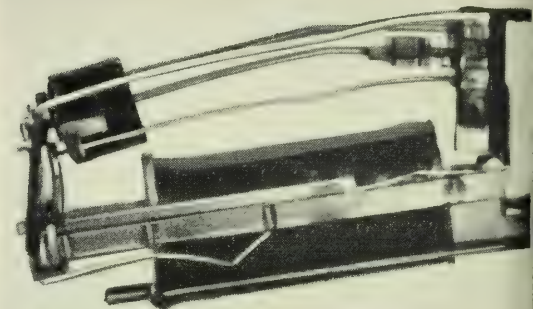
6

7

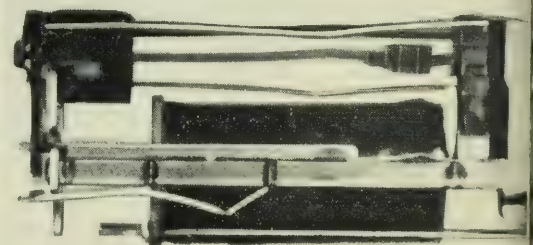
100



INSTANT OF IMPACT



MAXIMUM DISTORTION



COMPLETELY RESTORED

Fig. 15 — High-speed motion pictures of a falling relay at the moment of impact.

parts of the structure control the light falling on the cell, whose output can then be converted so as to read directly as displacement. The displacement data may then be differentiated as a function of time, in electrical differentiating circuits, to give very accurate measurements of velocity. A complete description of this method, originally due to E. L. Norton, has been published by M. A. Logan.⁵ As a result of the accuracy and convenience of this method, it has recently been used extensively in relay studies correlating the amount of contact chatter with armature velocity.

The motion of parts may also be observed with various forms of transducers. One which is now finding application in telephone relay studies is a type of electrostatic gauge, measuring armature displacement by the changes in capacity between it and a fixed electrode. The entire scheme is to be described by T. E. Davis and A. L. Blaha in a forthcoming issue.

Some relay motions need to be studied in three dimensions. Then the high-speed motion picture gives best results. By photographing at up to 5,000 frames per second, and viewing at about 20 frames, a time magnification of 250 to 1 can be gained. If need be, such pictures may be scaled off to give displacement-time information, a somewhat tedious task. A recent study of the effect of dropping in shipment gives a striking picture of the utility of the method. Fig. 15 gives views, photographed at 3,000 frames per second, of an AF relay as its mounting plate strikes a concrete block at the end of a six inch fall. Severe distortion, followed by recovery of the parts to normal, may be seen. With such information, relay designers can plan parts to stand the service stresses, and form a clear judgment of the margin of reliability built into the structure.

CONCLUSION

Although the relay is one of the oldest devices in the telephone business, many features of its operation are still imperfectly understood, even today. The number and complexity of the continuing technical problems may be judged from the other articles in this issue, on representative relay subjects. As improved relay operation becomes ever more important in the telephone system, the analytical and measuring technology for these devices must progress, in parallel. Some of the typical measuring equipment, as needed for modern relay design, has been described in this article.

REFERENCES

1. R. L. Peek, Jr., Measuring the Pull of Relays, Bell Lab. Record, June, 1953.
2. T. E. Davis, Measuring the Load-Displacement in Relays, Bell Lab. Record, June, 1953.

3. G. S. Burr, Servo-Controlled Tensile Strength Tester, *Electronics*, May, 1949, and E. G. Walsh, Continuously Recorded Relay Measurements, *Bell Lab. Record*, Jan., 1954.
4. P. P. Cioffi, Recording Fluxmeter of High Accuracy and Sensitivity, *Rev. of Sci. Instr.*, July, 1950.
5. R. L. Peek, Jr., Analysis of Measured Magnetic and Pull Characteristics, page 79 of this issue.
6. W. B. Ellwood, A New Magnetomotive Force Gauge, *Rev. Sci. Instr.*, **17**, p. 109, 1946.
7. W. P. Mason and S. D. White, New Techniques for Measuring Forces and Wear, *B. S. T. J.*, May, 1952.
8. M. A. Logan, Dynamic Measurements on Electromagnetic Devices, *B. S. T. J.*, Nov., 1953.

Magnetic Design of Relays

By R. L. PEEK, Jr., and H. N. WAGAR

(Manuscript received September 24, 1953)

The mechanical work and the speed of operation of telephone relays are determined in a large measure by the characteristics of the relay magnet. The underlying magnetic principles and the resulting design relationships for magnets are discussed in this article, which is in part a review of the background material and in part a description of its use in developing methods for magnet design and the analysis of magnet performance.

Basic energy considerations are shown to determine the relations between the work capacity and the magnetization characteristics, and analytical expressions for the latter are given in terms of the dimensions and materials of the magnet. These expressions are developed for the magnetic circuit approximation to static field theory, which is shown to provide an adequate representation of the field relations controlling performance. Methods are given for representing the magnetic circuit relations by means of a simple equivalent circuit. Expressions are derived for the mechanical output of the relay in parameters of this simple equivalent circuit, and these expressions used to determine optimum conditions for meeting specific design objectives.

INTRODUCTION

The complex switching equipment which handles the telephone traffic in automatic central offices is built up of simple component elements, of which the great majority are telephone relays. The large investment in these relays, of which tens of millions are made each year, has led to intensive effort to construct and use them as cheaply as possible, so that they will perform their function at a minimum over-all cost to the telephone system and thus to the subscriber. As the costs of use vary with the efficiency and speed, maximum economy requires the solution of technical problems in magnet design as well as the related problems of mechanical design for economy in manufacture. As a result, the technology of magnet design is under constant study at Bell Telephone Laboratories, directed to increased understanding of magnet performance

and to its improvement. This article gives the background of this technology as it applies to the dc telephone relay.

A telephone relay is an electromagnetic switch which actuates metallic contacts in response to signals applied to its coil. Its characteristics as a component of switching circuits are defined primarily by (1) the contact assembly, (2) the coil resistance, (3) the current flow requirements for operation, holding, and release, and (4) the operate and release times. The design of the contact assembly determines the number and nature of the contacts, the sequence of their operation, their current carrying capacity, and their reliability and life.

Structurally, the design of the relay, including both the electromagnet and the contact assembly, is governed not only by the performance requirements but also by the manufacturing considerations which determine the relay's initial cost, and by equipment considerations relating to the way it is mounted, wired, adjusted, and maintained. The ultimate objective is to perform a circuit function at a minimum over-all cost, including the costs of installation, power consumption, maintenance and replacement, as well as initial cost.

The design of the contact assembly determines a force-displacement characteristic representing the mechanical work which must be done by the electromagnet. The relations between this mechanical output and the electrical input to the coil are determined by the magnetic design of the relay, or specifically of its electromagnet. The electromagnetic design is therefore subject to the performance requirements, to the design of the contact assembly, and to the manufacturing and equipment considerations applying to the whole relay.

Magnetic design thus requires the ability to determine the effect on the performance of alternative choices of the configuration, dimensions, and materials of the electromagnet, of alternative load characteristics corresponding to variations in the design of the contact assembly, and of alternatives with respect to the coil dimensions and characteristics. The basic relations required for such prediction of performance are the magnetization relations, which express the field strength of the electromagnet as a function of the ampere turns, and the armature position.

The present article describes the evaluation of magnetization relations, and their use in relating the mechanical output with the electrical input to the coil. The time required for operation, which also depends upon the magnetization relations, is discussed in a separate article.¹ While much of the material given here has a broader application, the discussion of its use will be confined to the case of the simple neutral (non-polar) relay.

To illustrate the objectives of magnetic design, reference may be made to Fig. 1, which shows the load characteristic of the recently developed general purpose wire spring relay for a particular contact arrangement. The figure is taken from an article² describing this relay, which discusses the economic and other considerations which governed its design. Included in the figure are curves showing the pull exerted by this relay's magnet for various values of coil ampere turns. These pull curves determine the ampere turns required to operate a contact arrangement having a specific load characteristic, such as that shown. By means of the magnetization relations, these pull characteristics can be related to the design of the electromagnet.

The notation used in this article conforms to the list that is given on page 257.

1 THE COIL CONSTANT

The relay coil characteristics of interest in its use as a circuit component are its resistance and the current flow for operation. Subject to some qualifications as to available voltages and wire sizes, these may be summarized in a statement of the steady state power I^2R supplied in operation. As illustrated in Fig. 1, the coil quantity determined jointly by the load and the magnetic design is the ampere turn value NI . The power and the ampere turn value are related by the coil constant G_c or N^2/R (which equals $(NI)^2/(I^2R)$). This quantity is the equivalent single turn conductance of the coil, and is usually expressed in mhos. It is determined by the coil dimensions, as can be shown as follows:

Let A be the area of a cross-section of the coil as cut by a plane through the coil axis. Let m be the mean length of turn, the arithmetic mean of the lengths of the inner and outer turns. Then the coil volume S equals Am . If a is the cross-section of the wire, the number of turns N equals eA/a , where e is the copper efficiency, or fraction of the coil volume occupied by conductor. Substituting S/m for A , N equals $eS/(am)$. The wire length is Nm , and hence the resistance R equals $\rho Nm/a$, where ρ is the resistivity of the conductor. Hence N/R equals $a/(m\rho)$, and the coil constant is given by:

$$G_c = \frac{N^2}{R} = \frac{eS}{\rho m^2}. \quad (1)$$

The coil constant is thus independent of the wire size, except to the minor extent that the copper efficiency e decreases as the wire is made finer. With this qualification, and assuming copper wire to be used, the

coil constant is wholly determined by the coil dimensions and the type of insulation used. Thus the power required for relay operation depends upon (1) the load characteristic of the contact arrangement, (2) the magnetic characteristics which relate the pull to the ampere turns, and (3) the coil dimensions as defined by S/m^2 , which determines the coil constant N^2/R , and thus the relation between ampere turns and power input.

The external dimensions of relays are largely determined by the coil

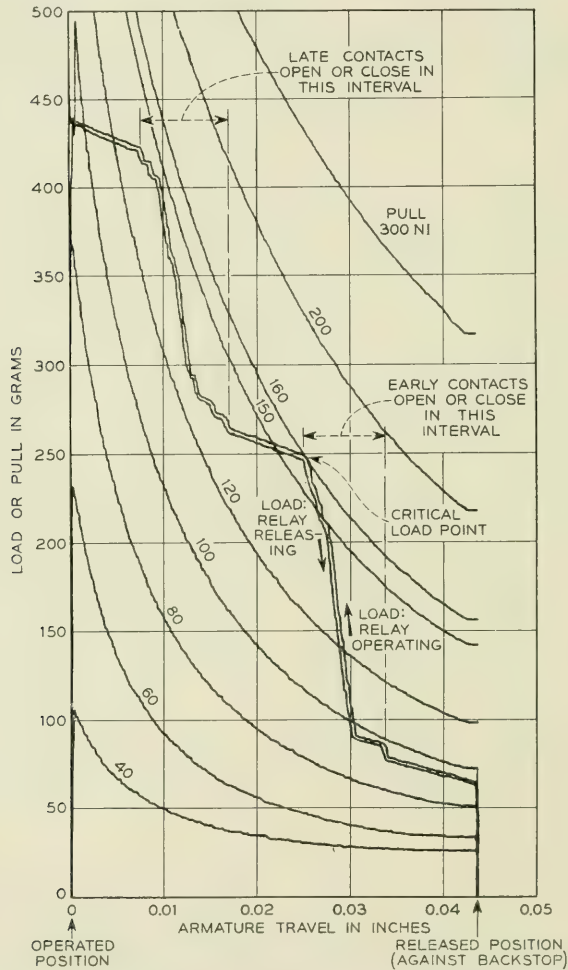


Fig. 1 — Typical load and pull characteristics of a wire spring relay with 12 transfer contacts.

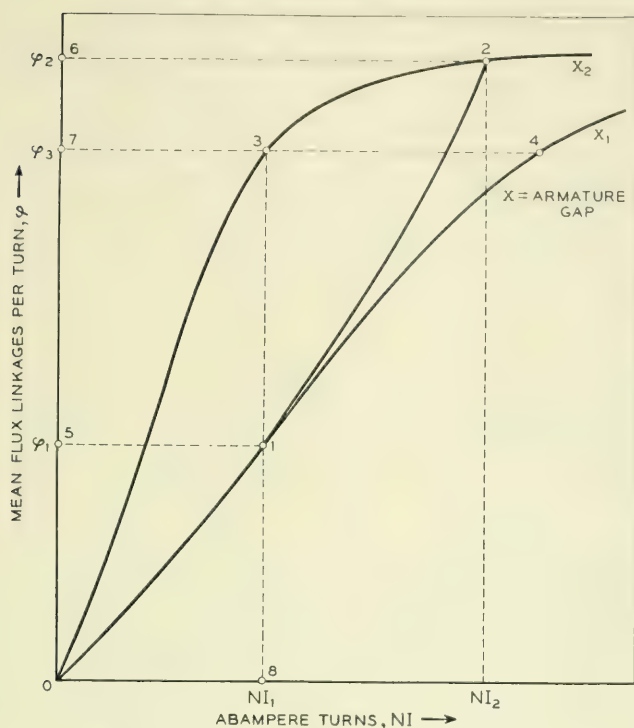


Fig. 2 — Energy relations in magnetization.

size, or winding space provided. The choice of these dimensions reflects an economic balance between the manufacturing costs of the magnet and its coil, which increase with these dimensions, and the savings in power cost resulting from increasing the coil constant.

2 MAGNETIZATION RELATIONS

The magnetization relations of an electromagnet define the steady state flux linkages of the coil as a function of the two variables: magnetomotive force $\mathfrak{F}$ (equal to $4\pi NI$), and the gap x , which specifies the armature position. They may be represented by a family of curves each giving the average flux linked per turn plotted against NI for a particular value of x . In Fig. 2, the curves marked x_1 and x_2 represent two such curves, where x_2 corresponds to a smaller gap than x_1 .

To determine how the magnetization relations depend upon the dimensions and configuration of the electromagnet requires their interpretation in terms of static field theory. Such interpretation is needed in deter-

mining the design conditions for attaining a desired performance. For a specific structure, however, the observed magnetization relations, apart from any other interpretation, provide a record of the part of the electrical energy input to the coil which is stored in the electromagnet. The performance of the magnet with respect to the mechanical work which this stored energy can do may be determined directly from the way in which this energy varies with armature position. The experimental determination of the flux φ for particular values of NI and x involves a measurement of the electrical quantity $N\varphi$ defined by the equation:

$$N\varphi = \int_{i=0}^{i=I} (E - iR) dt. \quad (2)$$

The electrical energy U stored in making this measurement is the time integral of $i(E - iR)$, or:

$$U = \int_{i=0}^{i=I} (E - iR)i dt = \int_0^{N\varphi} i d(N\varphi).$$

Hence a plot of $N\varphi$ versus I gives a measure of the stored energy U represented by the area between the curve and the axis of $N\varphi$. This conclusion is quite independent of any physical meaning attached to $N\varphi$ other than the definition of equation (2). Plotting φ versus NI , rather than $N\varphi$ versus I , is merely a change of scale, which does not affect the value of the area measuring the stored energy U . It is convenient to make this change of scale because the relation between φ and NI is independent of the number of turns N , provided the location and dimensions of the coil are unchanged. The preceding expression for U may therefore be written as:

$$U = \int_0^\varphi Ni d\varphi. \quad (3)$$

Thus for $NI = NI_1$ in Fig. 2, U is measured by the area 0-1-5 for $x = x_1$, and by the area 0-3-7 for $x = x_2$.

Magnetization curves have the general character shown in Fig. 2. They are approximately linear for small values of φ , but at higher values bend over and approach a limiting value, φ'' . This limiting value is the saturation flux, determined by the material and dimensions of the electromagnet, and designated throughout this paper by the double prime superscript.

The ratio of the magnetomotive force $\mathfrak{F}$ to the flux φ is the reluctance $\mathfrak{R}$, as defined by the equation:

$$\mathfrak{F} = 4\pi NI = \mathfrak{R}\varphi. \quad (4)$$

Over the linear portion of a magnetization curve, $\mathfrak{R}$ is a function of x only, or a constant for a particular curve. In this case, integration of equation (3) gives the following alternative expressions for the field energy U :

$$U = \frac{2\pi(NI)^2}{\mathfrak{R}} = \frac{\mathfrak{R}\varphi^2}{8\pi} = \frac{NI\varphi}{2}. \quad (5)$$

Over the upper curved portions of a magnetization curve, $\mathfrak{R}$ is a function of φ as well as of x , and equations (5) do not apply.

Decreasing Magnetization

In relay terminology, "operation," following closure of the coil circuit, is distinguished from "release," which follows opening of the circuit. The preceding discussion applies directly to the relations for increasing magnetization, as in operation. In release, the field energy and the current decrease together, giving a decrease in $N\varphi$ measured by a voltage time integral similar to the right-hand side of equation (2), but of opposite sign. The resulting decreasing magnetization curve is obtained by subtracting the decrease in $N\varphi$ from its initial value. The decreasing magnetization curve is higher than the magnetization curve, and the field energy recovered electrically is correspondingly less than that stored in magnetization; the difference corresponds to the loss of energy through hysteresis in the magnetic material.

Mechanical Output

Referring to Fig. 2, let the current have the steady value I_1 , with the armature at rest at x_1 . If the current increases to I_2 while the armature moves from x_1 to x_2 , the flux φ varies with Ni along some curve such as 1-2 corresponding to the values of x and Ni concurrently attained. The electrical energy drawn by the coil in this process (aside from the heating loss) is given by the integral of $i d(N\varphi)$, or $Ni d\varphi$, taken along the curve 1-2. Part of this energy appears in the increase in the field energy from U_1 to U_2 as given by equation (3) for the points 1 and 2 respectively. The balance represents the mechanical work done, the integral of Fdx from x_1 to x_2 where F is the pull. Hence:

$$\int_{x_1}^{x_2} F dx = \int_{\varphi_1}^{\varphi_2} Ni d\varphi - (U_2 - U_1). \quad (6)$$

The first right-hand term is represented in Fig. 2 by the area 5-1-2-6 while U_1 and U_2 are represented respectively by the areas 0-1-5 and

0-3-2-6. Thus the mechanical work, the left-hand term of equation (6), is represented by 0-1-2-3-0, the area bounded by the two magnetization curves and the path followed by φ , x and Ni in the concurrent change from I_1 at x_1 to I_2 at x_2 .

If armature motion occurs at constant flux, the first right-hand term in equation (6) is zero, and the mechanical work equals the change in the field energy U . If $\varphi = \varphi_3$ in Fig. 2 for example, the work done as the armature moves from x_1 to x_2 is $U_4 - U_3$, represented by the area 0-4-3-0. From equation (6), the pull F is then given by:

$$F = -\frac{\partial U}{\partial x} \quad (\varphi \text{ constant}). \quad (7)$$

If armature motion occurs at constant current, the first right-hand term in equation (6) becomes the change in $NI\varphi$. If $I = I_1$, for example, motion of the armature from x_1 to x_2 increases $NI\varphi$ from $NI_1\varphi_1$ to $NI_1\varphi_3$. Hence the mechanical work done is the difference between $NI_1\varphi_1 - U_3$, represented by the area 0-3-8 and $NI_1\varphi_1 - U_1$, represented by the area 0-1-8. The work done at constant current is therefore the change in the quantity W defined by the equation:

$$W = \int_0^{NI} \varphi d(NI),$$

which is represented by the area between the magnetization curve and the NI axis. From equation (6), the pull F is given by:

$$F = \frac{\partial W}{\partial x} \quad (I \text{ constant}). \quad (8)$$

The pull F can therefore be determined either from equation (7) or from equation (8), provided the magnetization curves are known. These equations may be applied graphically or numerically to compute the pull in specific cases. They may also be used, as shown in Section 7, to obtain expressions for the pull from expressions for the magnetization relations. In addition, they afford the following graphical interpretation of the dependence of the mechanical output upon the magnetization relations, the current, and the armature travel.

If x_1 in Fig. 2 represents the unoperated position of the armature, and x_2 its operated position, the work that can be done at constant current is $W_2 - W_1$, which varies with the value of I applying. For small values of I , for which both magnetization curves are linear, the work capacity $W_2 - W_1$ varies approximately as $(NI)^2$. At higher values of NI , the two magnetization curves approach each other as they approach the

limiting saturation flux, and the rate of increase of $W_2 - W_1$ becomes progressively smaller. The saturation flux therefore puts a ceiling on the mechanical work attainable.

If the armature travel is increased, increasing the unoperated gap x_1 , the corresponding magnetization curve is lowered, reducing W_1 with a consequent increase in the work capacity. However large x_1 may be made, there remains a leakage field, to which corresponds a limiting magnetization curve. An extreme upper limit to the work capacity is represented by the area between this limiting magnetization curve and that for the operated position of the armature.

The magnetization curves thus suffice for the evaluation of both the field energy and the mechanical output associated with armature motion, and therefore completely define the static performance of the electromagnet. The problem of relating this performance to the design then reduces to the problem of relating the magnetization characteristics to the design.

3 THE MAGNETIC CIRCUIT CONCEPT

A rigorous determination of the magnetization relations from the dimensions and configuration of the electromagnet would require the solution of the static magnetic field equations. For the geometry obtaining in actual structures, such solutions can at best be obtained only for specific cases, and then only by tedious numerical or graphical methods. Approximate solutions, however, may be obtained by a procedure in which the actual distributed field is taken as confined to a limited number of paths, which together form a network analogous to an electrical circuit. The extent to which this procedure may provide a valid approximation is indicated in the following brief review of the basic postulates of static field theory. For the present purpose, these may be stated as follows:

The energy of a static magnetic field is the volume integral of the product of the magnitudes of the B and H vectors over the space containing the field. These vectors coincide in direction at all points, and are subject to the conditions: (1) that B can be represented by closed continuous lines, ("lines of induction"), whose density measures the magnitude of B , (2) that the ratio μ of B to H at any point is determined by the medium in which the point is located, (3) that the line integral of H around a closed path is equal to 4π times the total current in all circuits linking this path. This integral is the magnetomotive force $\mathfrak{F}$; it has the value $4\pi NI$ for a coil of N turns with current I flowing in them.

For a rigorous general treatment, these postulates must be expressed in differential form (the field equations), and solutions obtained in which they are satisfied at all points within the field. They may be applied directly, however, in certain cases of simple symmetry, and this same treatment, to some measure of approximation, may be used more generally. To do this, the lines of induction measuring B can be considered as grouped into tubes of induction. Each such tube is, from the properties of B , continuous, and the integral of B over a cross section of the tube is a constant quantity φ , characterizing this tube. Over a length of tube $\Delta\ell$ bounded at its ends by two surfaces over each of which H is constant exists a difference in magnetic potential, $\Delta\mathfrak{F}$, equal to the line integral $\int H d\ell$. Then $\Delta\mathfrak{F} = \varphi\Delta\mathfrak{R}$ where $\Delta\mathfrak{R}$ is the reluctance of this portion of the tube, determined by its dimensions and the permeability μ of the medium. In particular, for a tube of uniform cross-section a over length ℓ , $\Delta\mathfrak{R} = \ell/(\mu a)$. More generally, if two equipotential surfaces can be identified, bounding a region of constant permeability (such as an air gap), the solution of Laplace's equation for the region bounded by these surfaces permits the evaluation of the flux between them, and thus of the reluctance $\Delta\mathfrak{R}$.

It follows that if the pattern of the field can be recognized, so that the boundaries of some major tube of force can be determined, the reluctance of its several sections can be evaluated, and their sum $\Sigma\Delta\mathfrak{R}$ or $\mathfrak{R}$, is the ratio $\mathfrak{F}/\varphi$, where φ is the flux of the tube and $\mathfrak{F}$ the magnetomotive force of the coil linking it.

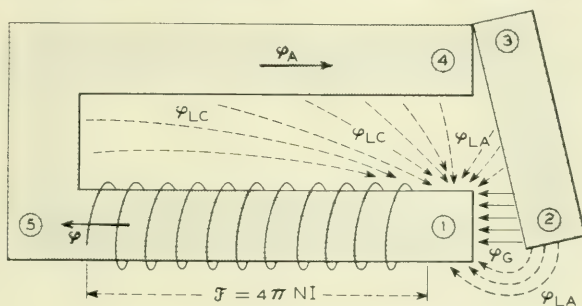
The possibility of recognizing the approximate pattern of the field results from the high values of the permeability of magnetic materials to that for air, μ_a . The ratio μ/μ_a is in excess of 1000 under normal conditions of operation. Hence the reluctances of air paths are large compared with those through magnetic material, the tubes of induction tend to follow a path through the iron, and the major changes in magnetic potential occur where they pass through air gaps. In an electro-magnet such as that of Fig. 3(a), the major tube of induction follows a path through the core and armature, and the potential drops balancing the applied magnetomotive force $\mathfrak{F}$ appear principally at the air gap separating surface 1 from surface 2, and at the heel gap separating surfaces 3 and 4. There will be little potential difference between 4 and 5, but the large potential difference between this region and surface 1 will result in a leakage field between them.

Thus the total flux φ linking the coil can be considered as divided into tubes of induction φ_A and φ_{LC} following the paths indicated in Fig. 3(a).

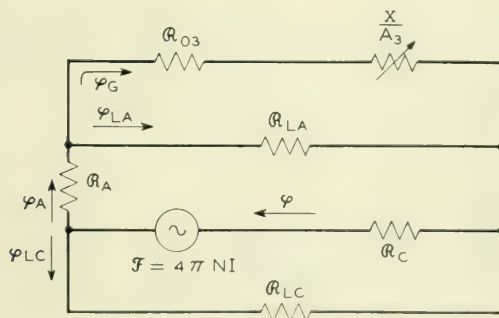
φ_A can be considered as subdivided in turn into $\varphi_{L,1}$ and φ_G representing respectively a leakage field from 2 to 1, and a field concentrated at the air gap between these surfaces.

These tubes of induction are analogous to the currents in an electrical network, and the applied magnetomotive force is analogous to the applied voltage producing these currents. The reluctances between equipotential surfaces are analogous to the resistances of the electrical network, and the potential differences between these surfaces are analogous to the voltage drops in the electrical resistances.

The field relationships involved may therefore be represented as in Fig. 3(b), in which the diagram of a resistance network indicates how the component reluctances determine the relation between the applied magnetomotive force $\mathcal{F}$ and the resultant flux φ . The potential difference between the points 1 and 5 at the ends of the coil is $\mathcal{F}$ less $\varphi \mathcal{R}_C$ where $\mathcal{R}_C$ is the reluctance of the core. This net potential balances the drop $\varphi_{LC} \mathcal{R}_{LC}$ across the leakage path and the drop through the armature path,



(a) MAGNETIC FIELD SCHEMATIC



(b) MAGNETIC CIRCUIT

Fig. 3 - Magnetic circuit representation of magnetic field relations.

comprising the drop $\varphi_A \mathfrak{R}_A$ through the iron part of this path in series with that across the two parallel paths followed by φ_{LA} and φ_G .

Obviously, this pattern can be elaborated and made more exact by more detailed consideration, particularly with respect to the leakage field across the coil, which is not wholly confined to that between its ends. In general, the field may be divided to any desired degree of refinement into tubes of induction, which may then be treated as a network. Thus, in principle, magnetic circuit representation may be used to evaluate the magnetization relations to any desired degree of accuracy.

Determination of the magnetization relations is thus reduced to the evaluation of the reluctances appearing in the magnetic circuit. These include both the reluctances of the iron parts for the major tubes of induction directed along the axes of these parts, and the reluctances of the air gaps and leakage paths. To evaluate these reluctances there are needed (1) values of the permeability of magnetic materials, as related to the flux density within them, and (2) expressions for the reluctances of gaps and leakage paths. These two topics are discussed in the sections following. While much of the material in these two sections is familiar, it is reviewed here with emphasis on the analytical formulation of the relations involved.

4 MAGNETIZATION CHARACTERISTICS OF MATERIALS

Magnetic properties are expressed either in terms of the relation between the permeability μ and the induction B , or of that between B and the potential gradient H , from which the $\mu-B$ relation is derived. The experimental determination of this relation is most commonly made with a uniformly wound ring sample, in which B and H are uniform throughout.

For increasing magnetization, as in relay operation, the pertinent $B-H$ relation is that obtained with an initially demagnetized specimen, shown as the solid line in the first quadrant in Fig. 4. This is the normal magnetization curve. It is nearly coincident with the locus of the terminal points (such as 1 and 2) of hysteresis loops obtained by cyclic magnetization and demagnetization.

The normal magnetization curve for magnetic iron is shown in Fig. 5 together with the corresponding $\mu-B$ curve. The permeability increases from its initial value μ_0 to a maximum value μ' at a density B' corresponding to the "knee" of the $B-H$ curve. (The single prime superscript is used throughout this paper to designate values of the various magnetic constants at maximum permeability.) Thereafter μ declines,

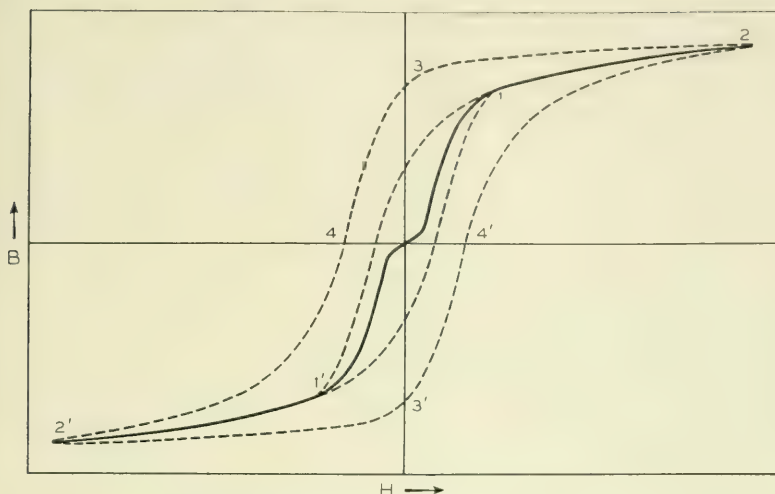


Fig. 4 — Cyclic magnetization relations.

and can be considered to approach zero as B approaches the saturation density B'' .

In the design and operation of ordinary electromagnets, the aspects of the $\mu-B$ relation of practical importance are (1) the order of magnitude of μ through the central portion of the curve, which determines the approximate magnitude of the (minor) contribution of the iron path to the total reluctance, and (2) the value of B at which μ becomes relatively small (less, for example, than 1000). At or near this value of B , the magnetic path reluctance increases rapidly for small increments in B , limiting the field strength and pull attainable. In the core, this limiting value of B , together with the cross section a , limits the flux $\varphi (=aB)$ which the core can supply; conversely, this value of B establishes the core cross-section a required to attain a desired value of φ . The exact value of B which is effectively limiting varies with the design conditions, but is approximately measured by the density B_M at which μ equals 1000.

Demagnetization Relations

For decreasing magnetization, as in relay release, the pertinent $B-H$ relation is the return portion of the major hysteresis loop, the loop starting from a point on the normal curve near saturation, such as the loop 2-3-4 of Fig. 4. Nearly the same loop is obtained for any location of the point 2 well beyond the knee of the curve. The distinguishing feature of this relation is the existence of the remanence B_R at the point 3, where

$H = 0$. To account for this there must be added to the magnetic postulates cited above the assumption that there may be a movement of electric charge within the material which supplies an effective mmf per unit length measured by the coercive force H_C , represented in Fig. 4 by the distance from the origin to the point 4.

When the coil circuit is opened in relay release, the coercive force of the core material results in a residual flux passing through the air gaps in series with the iron path. A potential drop $\mathfrak{F}$ must exist across these gaps. In the absence of applied magnetomotive force, this must be balanced by an equal and opposite drop through the core. There is thus an imposed negative potential gradient $-H$ per unit length of the core. The remanence B_R is therefore governed by the $B-H$ relation in the second quadrant, the curve 3-4 of Fig. 4. This is the demagnetization relation, as illustrated separately in Fig. 6. The intercepts of this curve on the B and H axes are H_C and B_R .

If $\mathcal{R}_E$ is the reluctance of the magnetic circuit external to the core, the external drop $\mathfrak{F}_E$ equals $\mathcal{R}_E\varphi$, where φ is the flux. Taking B and H as uniform in the core, $\varphi = Ba$ and $\mathfrak{F}$ equals $H\ell$, where a and ℓ are the cross-sectional area and length of the core. The values of B and H in the core must therefore conform to the equation:

$$-\frac{B}{H} = \frac{\ell}{a\mathcal{R}_E}. \quad (9)$$

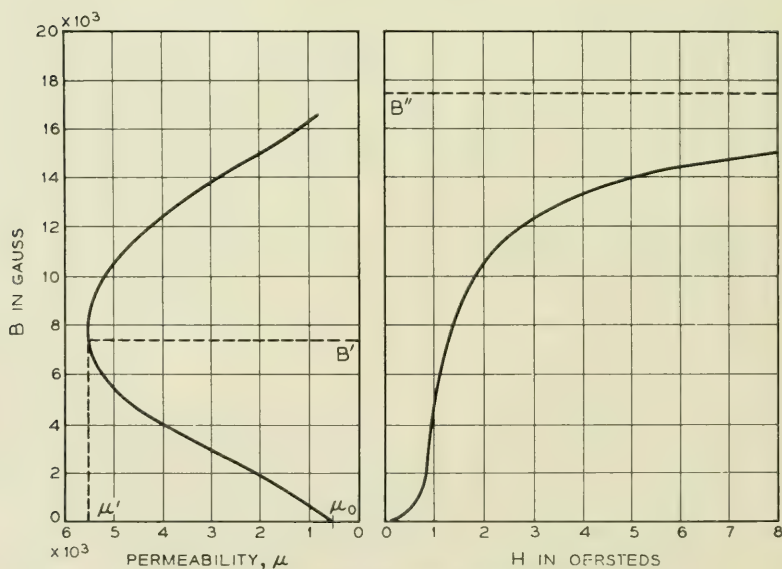


Fig. 5 — μ - B and B - H curves for magnetic iron.

As illustrated in Fig. 6, B and H are therefore determined by the intersection with the demagnetization curve of the line having a slope given by the right-hand side of (9). The complete magnetic circuit has a residual flux $\varphi = Ba$ resulting from the coercive mmf, $H_c \ell$, which equals the sum of the potential drop $H \ell (= \mathfrak{R} \varphi)$ external to the core and the drop $(H_c - H) \ell$ in the core.

Properties of Magnetic Materials

The properties of the magnetic materials commonly used for electromagnets are shown in Table I. This includes, in addition to the magnetic properties, the working properties in manufacture, and the resistivity. The latter determines the eddy current delay in otherwise similar structures, and is a controlling factor where fast release is desired.

For design purposes, allowance must be made for variations in the material, which are relatively large in the case of magnetic properties. Minimum and maximum values are therefore given in Table I. As indicated by the footnote in this table, H_c is subject not only to initial variation, but to an increase with time: "aging." The relatively large magnitude of this change for magnetic iron is a serious disadvantage in

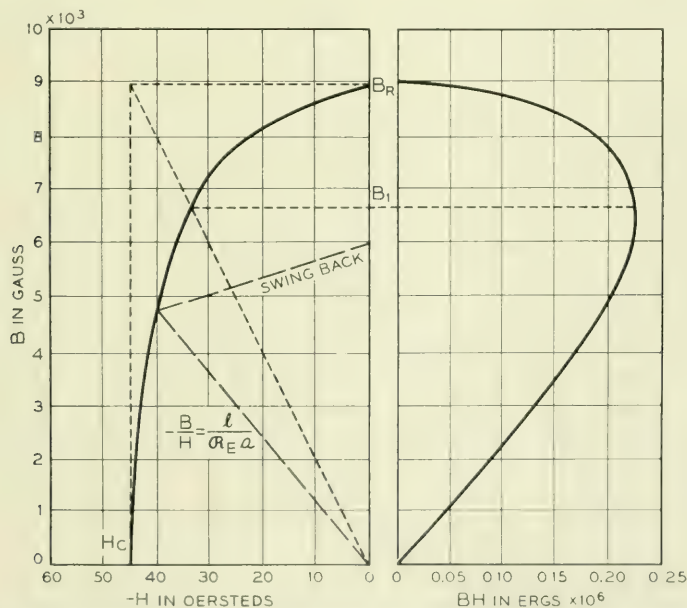


Fig. 6 Demagnetization and energy product curves ($1\frac{1}{2}$ per cent chrome steel).

TABLE I — SOFT MAGNETIC MATERIALS

Material	Magnetization				Demagnetization				Resistivity Microhms cm ⁻¹
	$\mu_{\max}$		B_M (gauss †)		B_R (gauss)		H_C (oersteds)		
	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.	
Magnetic iron									
Sheet	4,300*	12,000	15,500	16,200	10,500	15,000	0.5	1.4*	11
Rod	4,300*	8,000	15,500	16,200	10,500	12,000	0.5	1.4*	11
Mild steel	2,200	7,500	14,000	16,000	7,800	15,000	0.8	2.5	12
Cast iron	(Nominal: 600)		(B" = 10,000)		—		—		25
1% silicon iron	4,000	15,000	14,700	15,600	9,000	14,500	0.4	1.4	25
2½% silicon iron	4,000	12,000	14,000	15,000	8,000	12,000	0.4	1.4	40
4% silicon iron	5,000	12,000	13,500	15,000	8,000	12,000	0.3	1.1	60
H ₂ anneal iron									
Sheet	7,000	15,000	15,500	16,500	14,000	15,000	0.5	0.9	11
Rod	4,500	8,000	15,500	16,500	12,000	14,000	0.7	1.0	11
45 permalloy	15,000	60,000	14,000	15,000	8,000	12,000	0.1	0.4	50
78 permalloy	50,000	250,000	10,000	10,400	5,000	7,500	0.02	0.1	16
Nickel	(Nominal: 600)		(B" = 6,000)		—		—		8

* After aging, $\mu_{\max}$ for magnetic iron may be as low as 3000, and H_C as high as 2.5.

† B_M : Density for $\mu = 1000$. B": Saturation density.

applications in which release performance is important. Aging also reduces the permeability. In materials other than magnetic iron, the aging effect is relatively minor.

Because it is cheap and easily fabricated, magnetic iron has been the most commonly used magnetic material in Bell System relays and switching electromagnets. Other materials, particularly 45 permalloy, have been used in special applications where the improvement in performance, as in higher sensitivity or faster release, warranted the increased cost. The superiority of the permalloys in these respects is offset for heavy duty applications by the lower level of flux attainable with a given core cross-section, as measured by B_M .

The silicon steels are comparable with magnetic iron in cost, and similar in magnetic properties, except for slightly smaller values of B_M . They are, however, superior in their relative freedom from aging, and in their higher resistivity. These advantages are offset, in the case of the 4 per cent material, by its hardness and brittleness. All the silicon steels are used in sheet form for the construction of transformers and electrical machines. The 1 per cent and 2½ per cent materials are available in rod and bar form, and have working properties intermediate between those of magnetic iron and 4 per cent silicon iron. Because of its advantages in aging and resistivity, 1 per cent silicon iron has been used in preference

to magnetic iron in the recently developed wire spring general purpose relay.

Magnetic iron is a very low carbon steel of high purity. Commercial mild steel, properly annealed, will serve as an adequate substitute where the spread in properties shown in Table I can be tolerated.

In magnet design, casting offers the possibility of providing a more complex one-piece configuration than can be obtained with punched and formed parts. Casting is rarely used in current magnet construction

an illustrative exception is the British Post Office stepping magnet for their step-by-step switch. Grey cast iron has been a preferred material for this type of construction.

The properties of nickel are included because of its use as a magnetic separator and hinge member.

Hyperbolic Approximation to Magnetization Curves

The variation of permeability with density makes it necessary to provide some formulation of the μ versus B relation in developing an analytical treatment of magnetization relations. The B versus H relation for decreasing magnetization, the loop 2-3-4 of Fig. 4, has a shape similar to that of a rectangular hyperbola, asymptotic to the line representing B'' . This curve can therefore be represented approximately by the equation:

$$\frac{B}{B'' - B} = \frac{\mu''}{B''} (H_c - H), \quad (10)$$

in which μ'' , B'' , and H_c are constants.

This purely empirical relation is called the Froelich-Kennelly equation. In general, it does not give a satisfactory fit to the whole loop, but provides a satisfactory approximation for engineering use to the portions of the curve in either the first or second quadrants, using different values for the constants in the two cases. In addition, it may be employed to represent the upper portion of the normal, or increasing magnetization curve.

The expression for the permeability μ , or $B_c/(H_c + H)$ corresponding to (10) is:

$$\frac{1}{\mu} = \frac{1}{\mu''} + \frac{H_c + H}{B''}. \quad (11)$$

Hence, to the extent the B - H curve conforms to equation (10), the reciprocal of the permeability varies linearly with H . Alternately, the

permeability may be expressed in terms of B , giving the equation:

$$\mu = \frac{\mu''(B'' - B)}{B''}. \quad (12)$$

If this relation is applied to a part of a magnetic circuit, such as the core of a relay, the reluctance $\mathfrak{R}_c$ of this part may be written as $\ell/(\mu a)$, where ℓ is the length and a is the cross-sectional area of the part. Then from (10), $\mathfrak{R}_c$ is given by:

$$\mathfrak{R}_c = \mathfrak{R}_c'' \frac{\varphi''}{\varphi'' - \varphi}, \quad (13)$$

where $\mathfrak{R}_c'' = \ell/(\mu'' a)$, $\varphi = Ba$, the flux through the part, and φ'' is the saturation value of φ .

For increasing magnetization, this expression is applicable only for values of B , or φ/a , beyond the knee of the magnetization curve, the point of maximum permeability. Thus (13) is applicable for values of B above B' , the density at which μ has its maximum value μ' . Writing $\mathfrak{R}_c'$ for $\ell/(\mu' a)$ and φ' for $B'a$, (13) may be written in the alternative form:

$$\mathfrak{R}_c = \mathfrak{R}_c' \frac{\varphi'' - \varphi'}{\varphi'' - \varphi}. \quad (13A)$$

For values of B below B' the permeability of initially demagnetized material varies greatly with B , as shown in Fig. 5. In normal use, however, an electromagnet is rarely operated from a fully demagnetized state. Fig. 7(a) shows the magnetization relations for an electromagnet in repeated operation, and Fig. 7(b) shows the corresponding relations for its core. The solid line corresponds to initial operation from a demagnetized condition, the dotted line to decreasing magnetization in release, and the dashed line to subsequent remagnetization. The latter corresponds to higher permeability and a lower core reluctance than those for initial magnetization. As a convenient approximation, linear magnetization may be taken as a representative condition for B less than B' , in which case the core reluctance is constant for φ less than φ' . In this low density region, then, the magnetic circuit constants may be considered to be independent of the flux. This, of course, is never strictly true, but it is a satisfactory approximation for most ordinary electromagnets.

Hyperbolic Approximation to Decreasing Magnetization Curves

As the hyperbolic approximation is a purely empirical relation, it may be applied to the $\varphi - \mathfrak{F}$ relation for an electromagnet as well as to that

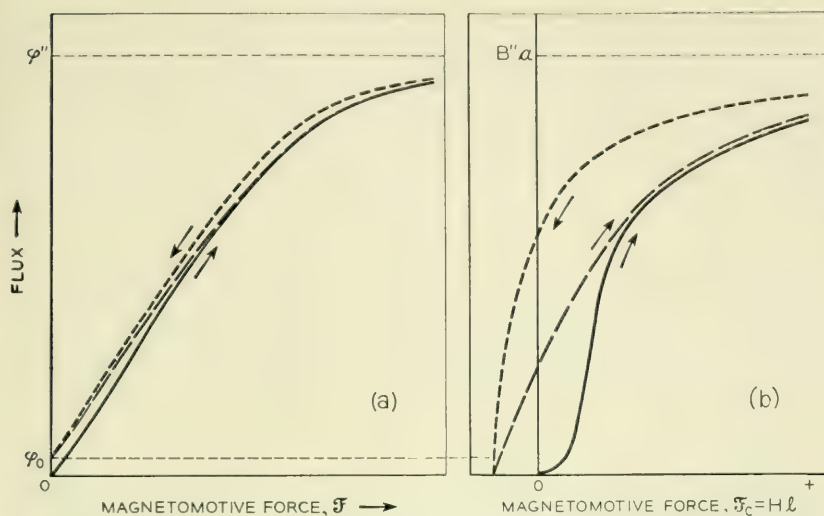


Fig. 7 — Repeated magnetization of an electromagnet and its core.

of a core or other part. It is convenient to use it in this way for the decreasing magnetization relation, as this is of interest only for a single value of the gap reluctance, that for the operated position.

The decreasing magnetization relations for an electromagnet and for its core are shown in Figs. 7(a) and 7(b) respectively. For decreasing magnetization, a residual flux φ_0 remains when $\mathfrak{F}$ is reduced to zero, determined by the same equilibrium conditions as apply to permanent magnets. The curve for the magnet, Fig. 7(a), must pass through φ_0 and be asymptotic to the saturation flux φ'' .

This relation between φ and $\mathfrak{F}$ may be represented by an equation of the same form as equation (10), with B and B'' replaced by φ and φ'' , with H and H_c replaced by $\mathfrak{F}$ and $\mathfrak{F}_c$, and the constant μ'' replaced by $1/\mathfrak{R}''$. Using the condition that $\varphi = \varphi_0$ for $\mathfrak{F} = 0$ to eliminate $\mathfrak{F}_c$, the equation may be written in the form

$$\frac{\mathfrak{F}}{\mathfrak{R}''\varphi''} = \frac{\varphi}{\varphi'' - \varphi_0} - \frac{\varphi_0}{\varphi'' - \varphi_0}. \quad (14)$$

If the length ℓ and cross-section a of the core are known, together with the magnetic constants of the core material and the reluctance $\mathfrak{R}$ of the return path (external to the core), the constant terms in (14) may be evaluated. φ'' is equal to aB'' , where B'' is the saturation density of the core material. As previously noted, the values of B_M in Table I may be used as effective values of B'' . φ_0 may be evaluated as the permanent

magnet flux supplied by the core to an external path of reluctance $\mathcal{R}$, using equation (9) and the demagnetization curve for the core material. $\mathcal{R}''$ may be determined from the initial slope $\mathcal{R}_i = d\mathcal{F}/d\varphi$ at $\mathcal{F} = 0$. By differentiation of equation (14), this initial slope is given by the equation:

$$\mathcal{R}_i = \left(\frac{\varphi''}{\varphi'' - \varphi_0} \right)^2 \mathcal{R}'' \quad (15)$$

$\mathcal{R}_i$ may be evaluated as the sum of the core reluctance and the external reluctance $\mathcal{R}$. In determining the core reluctance, μ should be taken as the incremental permeability, or the slope of the demagnetization curve at $B = \varphi_0/a$. With $\mathcal{R}_i$ thus evaluated, $\mathcal{R}''$ is given by equation (15).

5 MAGNETIC RELUCTANCES AND CONDUCTANCES

As shown in Section 3, determination of the magnetization or $\varphi - \mathcal{F}$ relations is equivalent to determining the reluctance $\mathcal{R}$, or $\mathcal{F}/\varphi$, and this in turn reduces to the determination of the component reluctances of the magnetic circuit. Some of the more useful expressions for the evaluation of these component reluctances are given in this section. Where parallel paths appear, the computations involve the reciprocals of reluctances. It is convenient to refer to the reciprocal of a reluctance as a magnetic conductance, or permeance.

Toroid

The magnetic field of a toroidal coil is shown in Fig. 8. Provided the medium within the coil is homogeneous, and the section diameter small compared with the toroid's diameter, symmetry requires the field to have the simple character shown. As $\mathcal{F} = H\ell$ and $B = \mu H$,

$$\mathcal{R} = \frac{\mathcal{F}}{\varphi} = \frac{H\ell}{\mu Ha} = \frac{\ell}{\mu a}, \quad (17)$$

which applies whether the path is in iron or air.

If the toroid is of iron, and the coil is concentrated over part of the length, some of the field appears outside the toroid. This effect is secondary, and (17) still applies to a close approximation. If a cut is made and an air gap introduced, the total reluctance is greatly increased, but the field within the toroid retains the same character except near the gap. The reluctance of the iron path, now in series with the air gap reluctance, is given by (17). This expression is therefore of general use in determining the reluctance of iron parts of length ℓ and uniform cross-section a .

Air Gap Between Parallel Planes

For magnetic paths in which the flux is uniform, it was shown above that the reluctance is given by (17). This applies to the case of an air gap between parallel planes of area A , as shown in Fig. 9, where the length of the path is the separation x . The reluctance of such a gap may therefore be written as:

$$\mathcal{R} = \frac{x}{A}.$$

In this, as in subsequent expressions for air gap reluctance, the multiplier $1/\mu_a$, which is the reciprocal of the permeability of air, is omitted for convenience. In CGS units, $\mu_a = 1$.

Special Gap Shapes

The simple relation just given provides a basis for the calculation of more complex shapes. For example, the reluctance of the wedge shaped gap of Fig. 10 may be found by assuming tubes of flux in parallel throughout the gap. Then the permeance of an elementary path is $\Delta\mathcal{P} = b\Delta r/r\theta$ and the total permeance of the gap is the sum of these elementary paths, or:

$$\mathcal{P} = \frac{b}{\theta} \int_{r_1}^{r_2} \frac{dr}{r} = \frac{b}{\theta} \ln \frac{r_2}{r_1}.$$

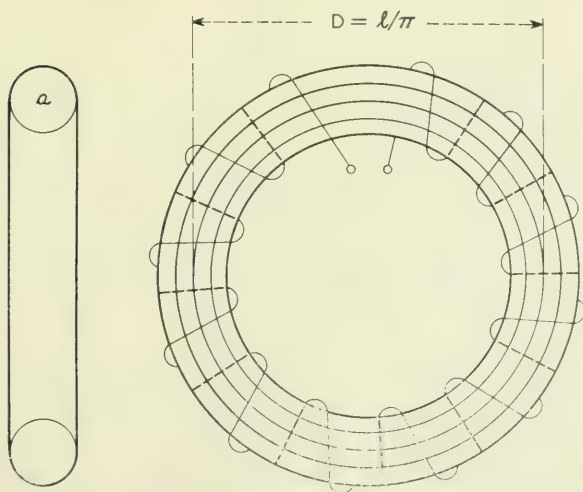


Fig. 8 — Field of a uniformly wound toroidal core.

The reluctance, as given in the figure, is the reciprocal of this.

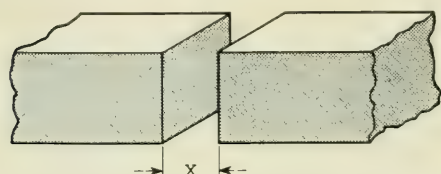
For the gap of Fig. 11, it is more convenient to estimate reluctances, summing the elementary series contributions to the entire gap. For this case:

$$\Delta \mathcal{R} = \frac{\Delta r}{br\theta},$$

where

$$\mathcal{R} = \frac{1}{b\theta} \int_{r_1}^{r_2} \frac{dr}{r} = \frac{1}{b\theta} \ln \frac{r_2}{r_1},$$

as noted in the figure.



$$\mathcal{R} = \frac{x}{A}$$

WHERE A = AREA OF ONE POLE FACE

IF THE TWO POLE FACE
AREAS ARE UNEQUAL,
USE THE SMALLER VALUE

Fig. 9 — Reluctance between parallel plane surfaces.

Effective Pole Face Area of an Electromagnet

As illustrated in Figs. 9, 10, and 11, an individual air gap has a reluctance represented approximately by an expression of the form:

$$\mathcal{R} = \frac{x}{A},$$

where x is the separation measured at the centroid of the area A . Armature motion is usually rotary, and a convenient point for the measurement of armature position may be at some distance from the axis other than that of the centroid. If x is so measured, then the separation at the centroid is kx , where k is the ratio of the lever arms, and hence the reluctance $\mathcal{R}$ is kx/A , equivalent to that of a gap of separation x and pole face area A/k . This area A/k is called the effective pole face area referred to the point at which x is measured.

In general, an armature has at least two gaps, as illustrated in Fig. 12. The expression for the armature reluctance must include the terms for both gaps which can conveniently be combined as follows. Let x be the armature motion measured at some distance ℓ from the axis of rotation, as indicated in the figure. Let k_1x be the corresponding separation at the centroid of a_1 , and k_2x the separation at the centroid of a_2 . Then the

reluctance of the combination is:

$$\mathfrak{R} = x \left(\frac{k_1}{a_1} + \frac{k_2}{a_2} \right).$$

It follows that the reluctance of the combined gap is equivalent to that of a simple gap with an effective pole face area given by:

$$\frac{1}{A} = \frac{k_1}{a_1} + \frac{k_2}{a_2}. \quad (18)$$

Exact: $\mathfrak{R} = \frac{\theta}{b \ln \left(\frac{1 + \frac{a}{2r_0}}{1 - \frac{a}{2r_0}} \right)}$

Approximate: $\mathfrak{R} = \frac{x}{ab}$

Errors of approximation:

Less than 1 per cent if $a < \frac{1}{3}r_0$

and/or $x < \frac{1}{2}r_0$

Less than 10 per cent if $a < r_0$

and/or $x < \frac{4}{3}r_0$

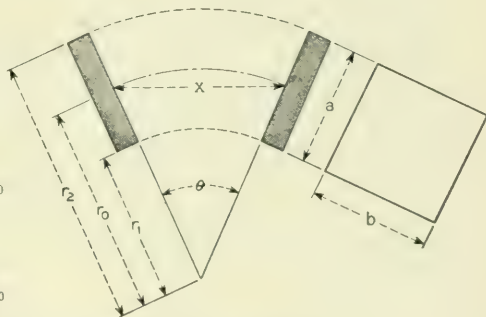


Fig. 10 — Reluctance between inclined plane surfaces.

Exact: $\mathfrak{R} = \frac{\ln \frac{r_2}{r_1}}{b\theta}$

Where b = length parallel to axis

Approximate: $\mathfrak{R} = \frac{x}{br_0\theta}$

Errors of approximation:

Less than 1 per cent if $r_2 < 1.4r_1$

Less than 10 per cent if $r_2 < 3r_1$

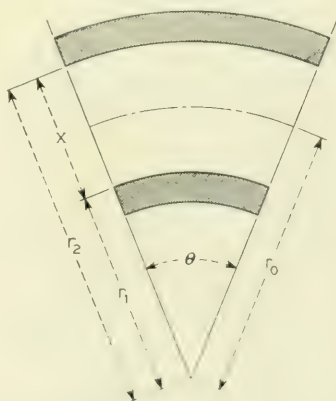


Fig. 11 — Reluctance of a cylindrical gap.

Leakage Reluctances

An accurate estimate of the reluctance between magnetic members requires a detailed knowledge of the flux paths. Since these are only known accurately in cases for which solutions of the field equations are available, approximations are obtained by assuming geometrical paths such as straight lines, arcs of circles, ellipses, and so forth. From these assumed paths the reluctance is calculated by means of the expression $\ell/\mu a$. The choice of suitable approximations depends largely upon a knowledge of the flux paths in certain simple cases which can be analyzed rigorously, and upon the experimental exploration of more complicated fields.

The method is satisfactory provided the separation between the magnetic members is small. It has been used to derive the relations given in Figs. 10 and 11. A further application of this method gives the reluctance between the side surfaces of coaxial cylinders, as shown in Fig. 13. This is useful in estimating the leakage reluctance shunting an air gap.

Where the separation between magnetic members is large, it is difficult to estimate the configuration of the flux paths. It is then necessary to employ the relations applying to the most nearly similar configuration for which a rigorous solution is known. Two such solutions are applicable to a number of problems. The first is the case of two infinitely long, parallel, equipotential, circular cylinders, shown in Fig. 14. The second is the case of two equipotential spheres of equal size, shown in Fig. 15.

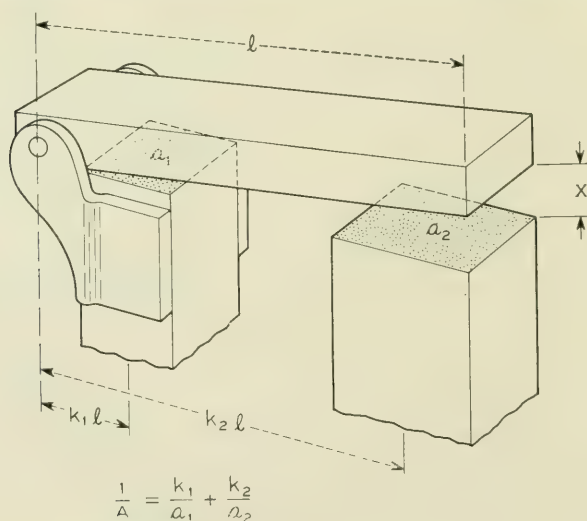
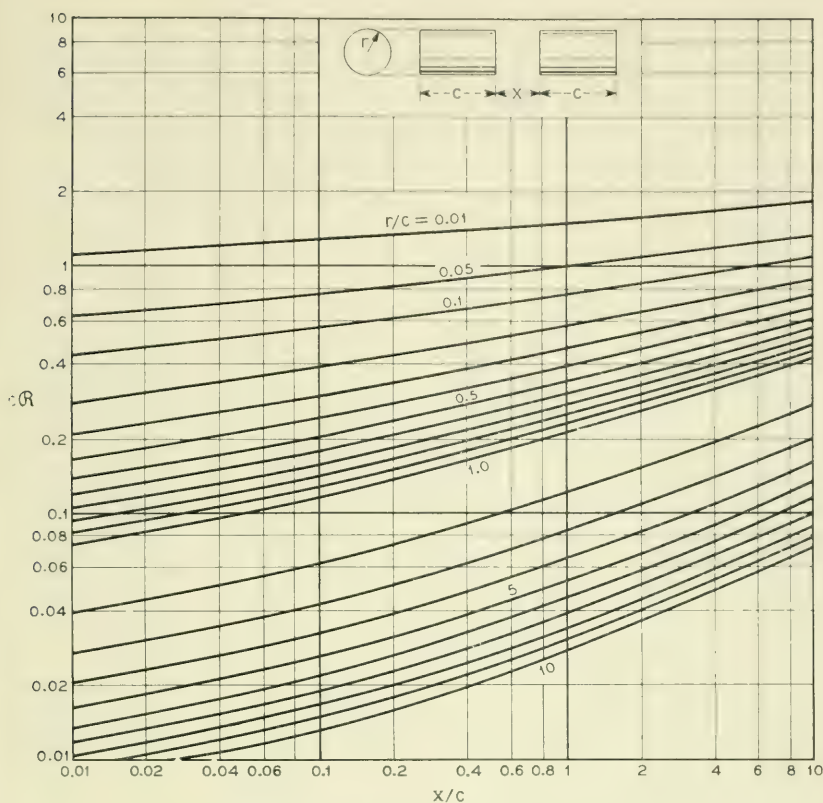


Fig. 12 — Effective pole face area referred to gap x at l .



$$m = \ell n \left[1 + 2 \frac{c}{x} \left(1 + \sqrt{1 + \frac{x}{c}} \right) \right],$$

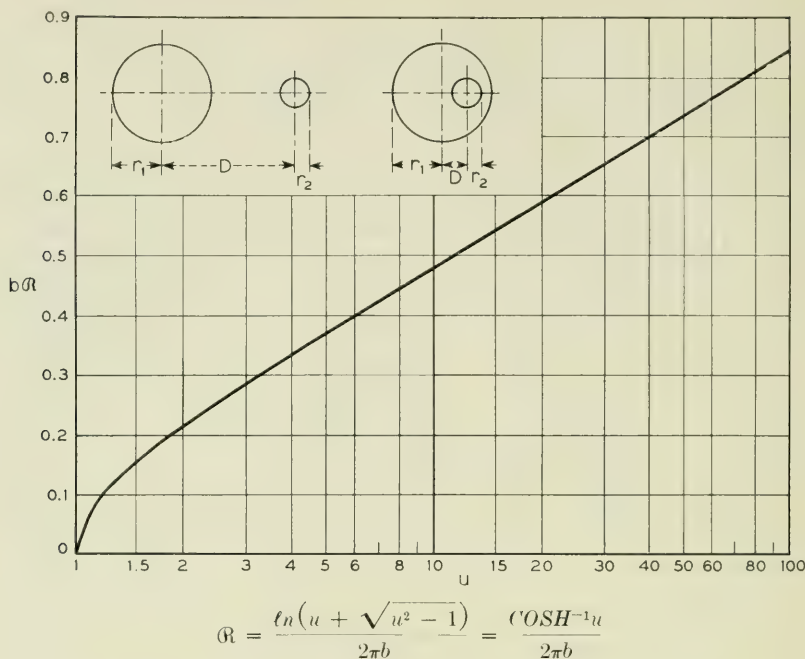
$$\text{when } \frac{c}{rm} < 1, \quad R = \frac{\cos^{-1} \left(\frac{c}{rm} \right)}{\pi rm \sqrt{1 - \left(\frac{c}{rm} \right)^2}},$$

$$\text{when } \frac{c}{rm} = 1, \quad R = \frac{1}{\pi c},$$

$$\text{when } \frac{c}{rm} > 1, \quad R = \frac{\ell n \left[\frac{c}{rm} + \sqrt{\left(\frac{c}{rm} \right)^2 - 1} \right]}{\pi rm \sqrt{\left(\frac{c}{rm} \right)^2 - 1}},$$

if cylinders are not circular, let $r = \frac{1}{2\pi} \times \text{perimeter}$.

Fig. 13 — Reluctance between side surfaces of end-on coaxial cylinders.



where $u = \frac{D^2 - r_1^2 - r_2^2}{2r_1r_2}$

and b = length of each cylinder

Fig. 14 — Reluctance between parallel cylinders.

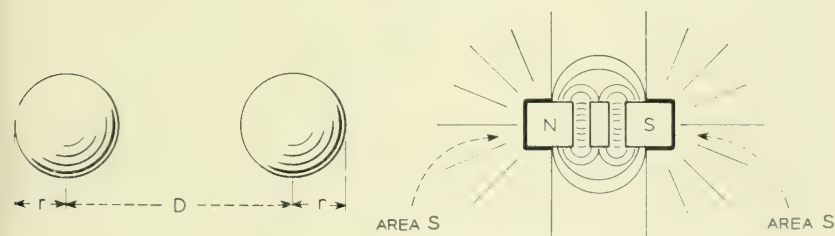
For other configurations which bear a reasonable resemblance to the above cases, the leakage reluctances can be estimated satisfactorily by judicious modifications of the expressions for the simple cases.

For example, consider the case of two parallel rectangular bars. They are roughly equivalent to two parallel circular cylinders provided the minimum separation is the same in both cases, and provided the perimeter of each cylinder is equal to the perimeter of the corresponding bar. Thus, to estimate the leakage reluctance between the rectangular bars, the radius of each equivalent cylinder is taken as $1/2\pi$ times the perimeter of the corresponding bar, and the center-to-center distance between the equivalent cylinders is taken as the minimum separation between the two bars plus the two equivalent radii.

Leakage between the legs of many magnet forms may be estimated by

the approximation just described. Consider the idealized form shown in Fig. 16, where the magnetic path consists mainly of two parallel cylinders connected at one end, the armature and working gap being at the opposite end. The leakage flux in parallel with the main gap flux is determined by the reluctance of the path between these cylinders. For that portion of the two cylinders appearing outside the coil and therefore at approximately constant potential, the reluctance is found from the relations given in Fig. 14. For the leakage reluctance over the length enclosed by the coil, the drop in potential along the core results in one-third the previous reluctance, as is shown in Section 6. The variation in permeance per unit length for such cases may be found in Fig. 17.

Leakage also occurs between the end sections of many magnet forms, and may be estimated by a procedure similar to those above, assuming the end surfaces to be equivalent to two hemispheres having the same diameter d as the cylinders. The quantity C_2d in Fig. 17 is one half the permeance between corresponding spheres, as determined from the relations of Fig. 15. From this the net leakage permeance of leg and end



$$\text{Exact: } \mathcal{R} = \frac{1}{2\pi r \left[1 + \left(\frac{r}{D}\right) + \left(\frac{r}{D}\right)^2 + \left(\frac{r}{D}\right)^3 + 2\left(\frac{r}{D}\right)^4 + 3\left(\frac{r}{D}\right)^5 + \dots \right]}$$

$$\text{Approximate: } \mathcal{R} = \frac{1 - \frac{r}{D}}{2\pi r}, \quad \text{when } D > 6r$$

$$\mathcal{R} = \frac{1}{2\pi r} = \frac{1}{\sqrt{\pi S}}, \quad \text{when } D \gg r$$

where S = surface area of one sphere.

Latter approximation may be used to estimate leakage reluctance between back surfaces of pole pieces.

Fig. 15 — Reluctance between spherical surfaces.

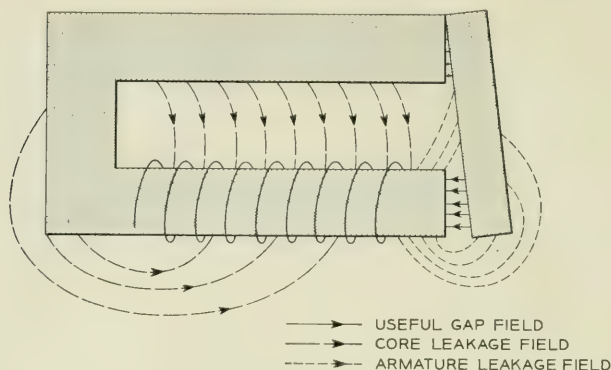


Fig. 16 — Distribution of the field of an electromagnet.

surfaces may be found as

$$\Phi = \frac{C_1 \ell_1}{3} + C_1 \ell_2 + C_2 d,$$

where values of C_1 and C_2 are given in curve form, and the equivalent value of d is as shown in Fig. 17. When the structure has two return paths, the reluctance is assumed to be one-half the value given by this last equation.

External Reluctance of a Bar Magnet

The relations of Figs. 13 through 17 suffice for the evaluation of leakage in most ordinary electromagnets. In special structures, particularly in polar relays, cases occur where the return path is predominantly an

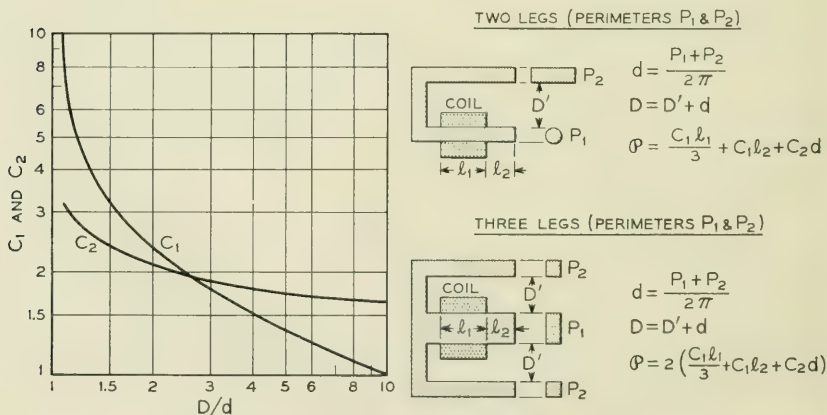


Fig. 17 — Effective reluctance between parallel bars joined at one end.

air path. A common case is that of a permanent magnet magnetized as a separate part prior to assembly in a polar structure. In such cases, the reluctance of the return path can be estimated as that of the external field of a bar magnet.

The reluctance of a bar magnet is closely represented by the reluctance of the same magnet in the form of a ring, in series with a reluctance representing all the flux return paths. Values for this reluctance may be assigned by measurements of magnetomotive force to produce a given flux in a ring sample and in a bar. The difference in magnetomotive force gives a measure of the reluctance of the air path. It has been found by

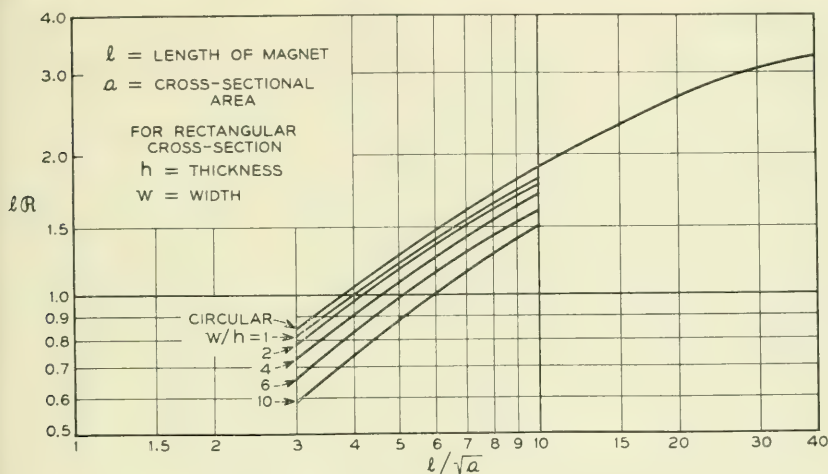


Fig. 18 — Effective leakage reluctance of a bar magnet.

Thompson and Moss⁵ that the reluctance per unit length of bar depends on the shape of the bar cross-section, and is a function of the magnet dimensions $l/\sqrt{a}$, as indicated in Fig. 18. The reluctance so shown includes all flux lines emanating from the bar, and so may be thought of equally as leakage or effective reluctance.

Cases such as the bar magnet are difficult to estimate because of the variable flux density along the length of magnetic material. Calculations have been found possible only for the case of an ellipsoid, which is rarely met in practice.

Reluctance of a Solenoid

Another special case is that of the air field of a coil, which has the character shown in Fig. 19. This type of field obtains in the initial flux

decay of an electromagnet when the core is saturated. A similar condition applies to the leakage field of enclosed reed relays where the magnetic material constitutes only a small part of the inside cross-section of the coil. In such cases the reluctance of interest is that characterizing the complete magnetic circuit of an air core solenoid.

A number of simple relationships for various coil configurations follow from the identities between reluctance and inductance. Inductance is defined as the ratio of flux linkages to applied current:

$$L = \frac{N\varphi}{I} = \frac{4\pi N^2 \varphi}{4\pi N I},$$

from which:

$$\mathfrak{R} = \frac{\mathfrak{F}}{\varphi} = \frac{4\pi}{\left(\frac{L}{N^2}\right)}, \quad (19)$$

showing the inverse relationship between $\mathfrak{R}$ and the single-turn inductance L/N^2 . Inductance for most of the usual coil configurations has been studied by Rosa and Grover⁷ who give relations from which reluctance may be estimated. Most such expressions are quite involved; among the simpler is that for the case where b and c in Fig. 19 are small compared to a , given by:

$$\mathfrak{R} = \frac{0.2317a + 0.44b + 0.39c}{a^2},$$

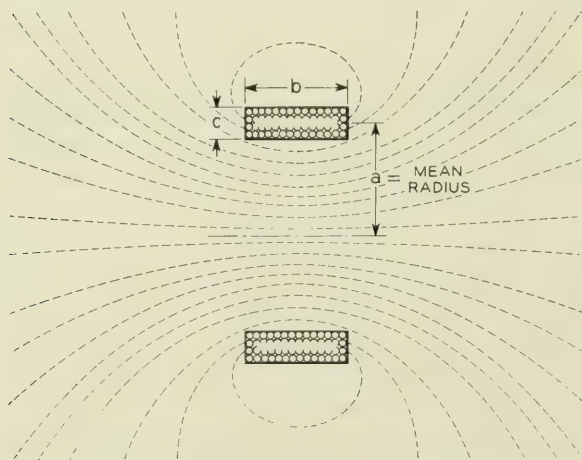


Fig. 19 Magnetic field of a solenoid.

and Maxwell's approximation for coil of rectangular section:

$$\alpha l = \frac{1}{a \left(\log \frac{8a}{R} - 2 \right)},$$

where R = geometric mean distance of coil cross-section.

The examples described in the present section cover those procedures which are used most frequently in estimating leakage reluctances. For a more extended treatment of this subject, reference may be made to S. Evershed³ and H. C. Roters.⁴

6 MAGNETIC CIRCUIT EVALUATION

In the discussion of the magnetic circuit concept in Section 3, it was noted that the accuracy of the representation varies with the extent to which the network of tubes of induction is sub-divided to correspond to the distributed nature of the actual field. The effect of sub-division upon the accuracy is greater in the high density region, where the reluctance of the iron parts is variable and increasing, than in the low density region, where the iron reluctance is small and approximately constant. Hence the choice of an adequate network is largely contingent upon the location of the iron parts having the highest flux density. Incipient saturation affects the reluctance of these parts, and thus the pattern of the field, while parts of lower density remain of low and approximately constant reluctance.

In ordinary electromagnets the core is the part of highest density, in which saturation limits the attainable field. It is good design practice to limit the core section to as small a value as is consistent with satisfactory performance, as this results in a minimum inside coil diameter. This is advantageous with respect to the coil constant, as shown by equation (1).

In the special case of high speed relays, it is advantageous to minimize the mass of the armature and hence its cross-section. In such relays armature saturation may control, or occur concurrently with core saturation. Saturation elsewhere than in the core or armature is of interest only in the diagnosis of faulty design, since the return members should have a section adequate to carry the maximum field at densities well below saturation.

Thus in most electromagnets, saturation occurs in the core, and incipient saturation affects its reluctance and the pattern of the associated field. The magnetomotive force varies along the length of the coil, and the core is therefore subject to variations along its length in magnetic

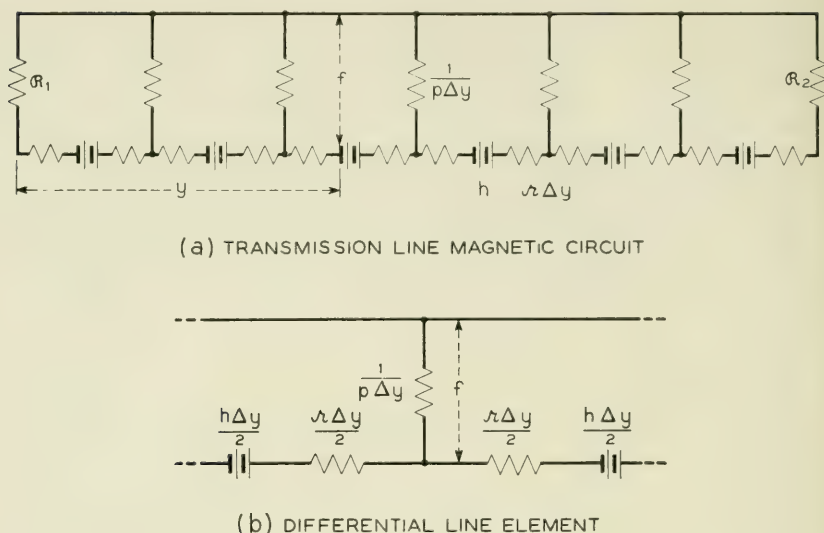


Fig. 20 — Transmission line analogy to the field of an electromagnet.

potential, in flux density, and (at high densities) in reluctance per unit length. To treat it as a single circuit element, as in the discussion of Fig. 3, is thus a highly simplified approximation. An understanding of the extent to which this approximation is valid may be gained by considering a more rigorous analysis, in which, as indicated in Fig. 20, the core and return path are treated as analogous to a transmission line with distributed constants.

Transmission Line Analogy

In Fig. 20(a), the core is shown as one side of the line, and the return path as the other side. The length of the line is taken as the length of the winding, which is assumed to be distributed uniformly. It is assumed that the line is terminated at its ends by lumped reluctances, R_1 and R_2 . The characteristics of the line and the applied magnetomotive force are expressed in terms of the following quantities:

f = mmf difference between the two sides of the line,

$\mathcal{L}$ = reluctance per unit length of line,

p = leakage permeance per unit length of line,

h = impressed mmf per unit length of line.

Instead of trying to represent the entire magnetic circuit by a simple network of a few lumped reluctances, it is assumed only that an infinitesimal length (dy) of the magnetic line can be represented by the "T"

network shown in Fig. 20(b). The complete magnetic circuit, then, consists of an infinite number of these elementary networks connected end-to-end and terminated in lumped reluctances at the extreme ends. Note that this approximation to the magnetic circuit explicitly recognizes the distributed nature of the leakage flux.

Consider now the several quantities which enter into the approximation. Each of the terminating reluctances ($\mathfrak{R}_1$ and $\mathfrak{R}_2$) includes two component reluctances in parallel. The first component is simply the leakage reluctance across the end of the line. The second component is the sum of the reluctances of the magnetic members and series air gaps which complete the circuit from one side of the line to the other. For any single value of the working gap, $\mathfrak{R}_1$ and $\mathfrak{R}_2$ may be taken as constants.

The quantity p , which is the leakage permeance per unit length between the two sides of the line, depends upon the geometry of the structure, and may be calculated by the methods of Section 5. Provided the configuration of the magnetic line is uniform throughout its length, p is taken as a constant. The assumption that p is constant is equivalent to assuming that all leakage paths between the two sides of the transmission line lie in planes that are perpendicular to the core. This condition is not satisfied near the ends of the core, but correction for the end effects may be made in evaluating the terminal reluctances, $\mathfrak{R}_1$ and $\mathfrak{R}_2$.

The quantity $\mathfrak{L}$, the series reluctance per unit length of line, involves magnetic material whose permeability varies with flux density. It is a variable whose magnitude depends upon the applied magnetomotive force, upon the terminating reluctances, and upon position along the line, since it is a function of φ . In the low density region, however, its value is substantially independent of φ . Application of the magnetic circuit relations results in the following equations:

$$\frac{df}{dy} = h - \mathfrak{L}\varphi, \quad (20A)$$

$$\frac{d\varphi}{dy} = -p\varphi, \quad (20B)$$

whence:

$$\frac{d^2\varphi}{dy^2} = -p(h - \mathfrak{L}\varphi). \quad (21)$$

Subject to the validity of the original assumptions, the solution of equation (21) describes the way in which the flux φ varies with y , the distance along the line.

At low densities, where $\mathcal{Z}$ is constant, (21) may be solved to give a picture of the flux distribution in the magnet. The solution is obtained in terms of hyperbolic functions. The resulting expression is somewhat cumbersome, but reduces to a much simpler form in the special case where $\mathcal{Z}$ is small compared with h . Neglecting $\mathcal{Z}\varphi$, equation (20A) reduces to:

$$\frac{df}{dy} = h \quad \text{or} \quad f = hy$$

In this case then, the magnetomotive force varies linearly along the core, as was assumed in the discussion of Fig. 17 in Section 5.

For $f = hy$, equation (20B) becomes:

$$\frac{d\varphi}{dy} = -phy,$$

whence:

$$\varphi = \varphi_0 - \frac{phy^2}{2}, \quad (22)$$

where φ_0 is flux at $y = 0$ (one end), and ph is a constant for a given magnet structure. Thus, for this approximate case, the core flux falls off along its length approximately as the square of the distance from one end. The second term in (22) represents the leakage flux. For the whole core, for which $y = \ell$, the leakage flux is given by $p h \ell^2/2$, or by $p \ell \mathfrak{F}/2$. The average flux linked per turn, however, is the integral

$$\frac{1}{\ell} \int_0^\ell \frac{phy^2}{2} dy,$$

corresponding to a leakage reluctance of $p\ell/3$. Hence the factor 3 is used to evaluate the average flux linked per turn in Fig. 17.

From this consideration of the more rigorous transmission line treatment, it is apparent that $\mathcal{Z}\varphi$ must be small compared with h for the lumped core and leakage reluctance approximation to be valid. If $\mathcal{Z}\varphi/h$ is small, the potential drop in the core, $\mathcal{R}_c\varphi$, is small compared to the applied magnetomotive force $\mathfrak{F}$. In most ordinary electromagnets, the ratio $\mathcal{R}_c/\mathfrak{F}$ is small in the low density region. For sensitive relays with long cores of small cross-section, this is not the case, and the lumped constant treatment is correspondingly limited in accuracy, even at low densities.

At high densities, however, the core reluctance increases even in ordinary electromagnets. To use the transmission line analogy at high densities, it is necessary to express $\mathcal{Z}\varphi$ in (20A) in terms of the Froehlich-

Kennelly relation described in Section 4. A solution to the resulting equations can be obtained in series form, but this solution is too complex for convenient use in engineering estimates. From this formulation of the problem, however, it is apparent that the use of lumped values of core and leakage reluctances at high densities can only be a rough approximation, as the reluctance per unit length must vary along the core, and the pattern of the leakage field and hence the leakage reluctance are no longer constant. However, in representing the core reluctance as increasing from its low density value and approaching infinity as $\varphi \rightarrow \varphi''$, the lumped approximation correctly represents the limiting conditions. It therefore provides a rough approximation to the intermediate values.

Series-Parallel Magnetic Circuit

Subject to the limitations discussed above, the magnetic circuit of most ordinary electromagnets can be represented in the form shown in Fig. 21. The core reluctance $\mathfrak{R}_c$ is in series with two parallel paths: a leakage path of reluctance $\mathfrak{R}_{L2}$, and an armature path of reluctance $\mathfrak{R}_{02} + x/A_2$. The subscript 2 is used with the constants of this particular circuit to distinguish them from those of the simpler approximation to be discussed below. The magnetic circuit of Fig. 3 reduces to that of Fig. 21 if the reluctance $\mathfrak{R}_A$ of the former is ignored, or considered as part of the reluctance $\mathfrak{R}_{02}$.

The circuit reluctance $\mathfrak{R}$, or $\mathfrak{F}/\varphi$, can be derived from the circuit by the procedure applying to resistances in an electrical circuit, and is given by:

$$\mathfrak{R} = \mathfrak{R}_c + \frac{\mathfrak{R}_{L2} \left(\mathfrak{R}_{02} + \frac{x}{A_2} \right)}{\mathfrak{R}_{L2} + \mathfrak{R}_{02} + \frac{x}{A_2}}. \quad (23)$$

The evaluation of the constants of this circuit may be described with

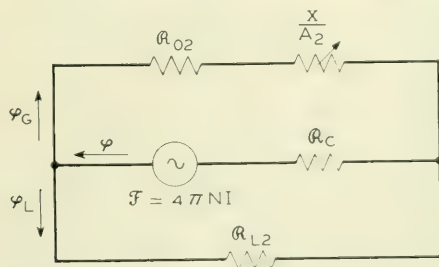


Fig. 21 — Series parallel magnetic circuit — the usual design analogy.

reference to the two structures shown schematically in Fig. 22, where Fig. 22(a) represents the familiar "end-on" armature type of construction, and Fig. 22(b) represents the "flat type" relay of Bell System use. The dashed lines indicate the path along which the length of the parts is measured, while the lined areas are those of the main, heel, and side gaps.

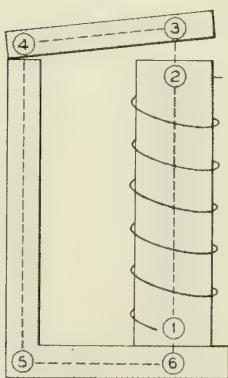
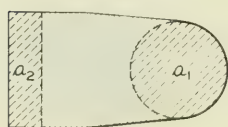
The core reluctance, $\mathcal{R}_C$, is determined from equation (17), taking the length as between the points 1 and 2 in both figures. The permeability is taken at the nominal maximum value throughout all iron parts.

The closed gap reluctance, $\mathcal{R}_{02}$, is the sum of the following:

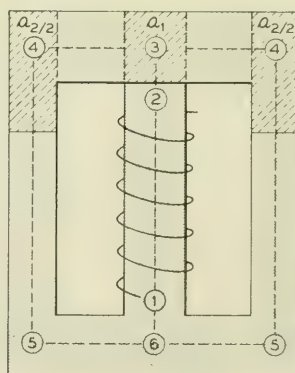
A. The iron reluctance, determined from equation (17) taking the lengths of the several parts as measured around the path 2-3-4-5-6-1. In each term the cross-sectional area a is that of the part through which the flux passes. In Fig. 22(b) of course, the two side sections are added to give the total section.

B. The contact gap reluctances computed as x/a_1 and x/a_2 , taking x as for an air gap of 0.005 cm. The reluctance of the joint at 1 in Fig. 22(a) is computed in the same way, and included in the sum.

C. The "stop pin" air gap reluctance, computed as x/A_2 , where A_2 is the effective pole face area as determined below, and x is the separation at the measuring point for the stop pin opening.



(a) END-ON ARMATURE TYPE



(b) FLAT TYPE

Fig. 22 — Magnetic circuit components of typical relay structures.

The effective pole face area, A_2 , is determined by equation (18), following the procedure described in the discussion of Fig. 12.

The leakage reluctance, $\mathcal{R}_{L2}$, is determined in the case of Fig. 22(a) by the procedure discussed in Section 5 and indicated in Fig. 17. In using this, ℓ_1 is the length 1-2, while ℓ_2 is the length 2-3 in Fig. 22(b) and $\ell_2 = 0$ in Fig. 22(a). The reluctance terms $C_1\ell_2$ and C_2d correspond to the armature leakage reluctance $\mathcal{R}_{LA}$ of Fig. 3, here taken as in parallel with the core leakage reluctance in determining $\mathcal{R}_{L2}$.

It should be noted that this procedure provides for two flux paths across each gap: the flux through the reluctance x/A_2 , varying linearly with gap, and the parallel leakage flux through a reluctance calculated as though the armature were absent. This representation allows for the effect of fringing, taking the total field across the gap as the sum of these two fields. It carries the implication that experimentally the two fields cannot be separated by search coil measurements.

For the magnetic circuit of Fig. 21, the low density reluctance is given by equation (23). The reluctance terms are calculated for maximum permeability μ' , corresponding to density B' , and hence for a total core flux $\varphi' = B'a$. As discussed in Section 4, these values are approximately applicable through the low density region, or for φ less than φ' . For φ greater than φ' , $\mathcal{R}_C$ is taken as given by equation (13). Thus in the high density region, the total reluctance may be written as:

$$\mathcal{R} = \mathcal{R}_C + \mathcal{R}_E,$$

where,

$$\mathcal{R}_E = \frac{\mathcal{R}_L \left(\mathcal{R}_0 + \frac{x}{A} \right)}{\mathcal{R}_0 + \mathcal{R}_L + \frac{x}{A}},$$

and,

$$\mathcal{R}_C = \mathcal{R}_C'' \frac{\varphi''}{\varphi'' - \varphi'},$$

in which,

$$\mathcal{R}_C'' = \frac{\ell}{\mu' a} \cdot \frac{\varphi'' - \varphi'}{\varphi''},$$

and $\varphi'' = B''a$, where ℓ and a are the length and cross-section of the core, respectively. The values of B_M in Table I may be taken as estimates of the effective value of the saturation density B'' .

Equivalent Magnetic Circuit

Over the low density region, in which the magnetic circuit reluctances are substantially independent of the flux, the expressions for the reluctance can be simplified by a procedure analogous to that used in deriving electrical network equivalents. Provided the gap reluctance varies linearly with the gap, magnetic circuits such as those of Figs. 3 and 21 can be replaced by an equivalent circuit of the simple form shown in

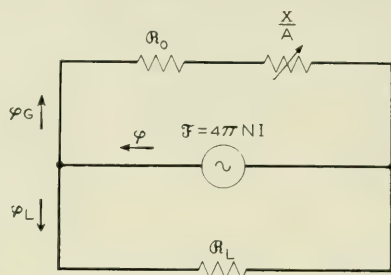


Fig. 23 — Equivalent magnetic circuit.

Fig. 23. For this simple parallel circuit, the total reluctance is given by:

$$\mathcal{R} = \frac{\mathcal{R}_L \left(\mathcal{R}_0 + \frac{x}{A} \right)}{\mathcal{R}_0 + \mathcal{R}_L + \frac{x}{A}}. \quad (24)$$

The simpler subscripts of these *equivalent* values of the magnetic circuit constants are used to distinguish them from the *design* values, applying to the magnetic circuit taken as representing the actual structure. In the usual case the design values apply to the two mesh circuit of Fig. 21, for which the additional subscript 2 is used. When a three mesh circuit is required to represent the structure, the design values are distinguished by the subscript 3, as in the constants of Fig. 3.

In the usual case in which the *design* values apply to the circuit of Fig. 21, expressions for the *equivalent* values may be obtained by comparison of the reluctance given by (23) with that given by (24). The former equation may be written in the form.

$$\mathcal{R} = \mathcal{R}_C + \mathcal{R}_{L2} - \frac{A_2 \mathcal{R}_{L2}^2}{A_2 \mathcal{R}_{02} + A_2 \mathcal{R}_{L2} + x},$$

while (24) may be written.

$$\mathcal{R} = \mathcal{R}_L - \frac{A \mathcal{R}_L^2}{A \mathcal{R}_0 + A \mathcal{R}_L + x}.$$

By comparison, these two expressions are identical for all values of x , provided the following conditions are satisfied:

$$\begin{aligned}\mathfrak{R}_L &= \mathfrak{R}_C + \mathfrak{R}_{L2}, \\ A\mathfrak{R}_L^2 &= A_2\mathfrak{R}_{L2}^2, \\ A(\mathfrak{R}_0 + \mathfrak{R}_L) &= A_2(\mathfrak{R}_{02} + \mathfrak{R}_{L2}).\end{aligned}$$

Upon collecting terms, the relations between design and equivalent circuits are given by:

$$\begin{aligned}\mathfrak{R}_L &= \mathfrak{R}_C + \mathfrak{R}_{L2}, \\ A &= A_2/p^2, \\ \mathfrak{R}_0 &= p^2\mathfrak{R}_{02} + p\mathfrak{R}_C,\end{aligned}\tag{25}$$

where:

$$p = 1 + \mathfrak{R}_C/\mathfrak{R}_{L2}.$$

Thus the magnetization relations in the low density region can be represented by the reluctance given by (24), corresponding to the simple parallel circuit of Fig. 23, provided the constants are evaluated from those of the design circuit by means of equations (25). The evaluation of the pull and of the field energy in the low density region can therefore be conducted in terms of the simpler relations applying to the equivalent circuit.

As these simpler relations suffice to define the performance, they may be readily evaluated experimentally, using the procedures described in a companion article.⁸ While these equivalent constants provide a simple and convenient means for the description and analysis of performance, they are related to the dimensions of the structure only through the conditions of equivalence given by equations (25).

Special Magnetic Circuits

The reluctance of most ordinary electromagnets can be expressed in terms of the series parallel magnetic circuit of Fig. 21, as in the two cases of Fig. 22 discussed above. Differences in configuration may affect the detailed procedure for estimating the constants, but the same circuit schematic applies. There are, however, other magnetic structures requiring different magnetic circuits for their representation. For purposes of illustration, a summary discussion of three such cases is given here.

For *armature saturation*, or cases where the flux density in the armature exceeds that of the core, as in some high speed relays, the magnetic cir-

cuit must be taken as that of Fig. 3. In this case it is the armature flux φ_A , rather than the total flux φ , which approaches a saturation value limiting the mechanical output. In the low density region, where the reluctances are constant, the conditions of equivalence given by equations (25) may be used to evaluate a simple parallel equivalent to the armature path reluctance, reducing the circuit to the form of Fig. 21, and this may in turn be reduced to an equivalent circuit of the form of Fig. 23. In the high density region, however, it is the armature rather than the core reluctance which must be expressed in terms of the Froelich-Kennelly approximation.

For *reed relays*,⁶ the structural schematic is as shown in Fig. 24. In some cases, the external shield may be omitted. The magnetic circuit involves an air path through the coil in parallel with a path through the reeds. With a new interpretation of the terms, the circuit of Fig. 3 may be taken as applying, with $\mathcal{R}_C$ taken as zero, and $\mathcal{R}_{LC}$ representing the air leakage field. With some correction for the effect of the shield, this may be estimated by means of the solenoid reluctance expressions of Section 5. The controlling path is that through the reeds, represented by $\mathcal{R}_A$ for the saturating section of maximum density in series with the parallel paths between the reeds: the useful path at the gap and the leakage field. The latter may be estimated from the relations of Fig. 13.

For *polar relays*, the network contains both the permanent magnet magnetomotive force $\mathcal{F}_M$ and the coil magnetomotive force $\mathcal{F}_C$. A typical case is shown in Fig. 25. The constants applying can be evaluated by the procedures described in Section 5, and the somewhat complex circuit equations resulting can be written by the application of Kirchhoff's laws in a manner wholly analogous to that for the corresponding electrical circuit.

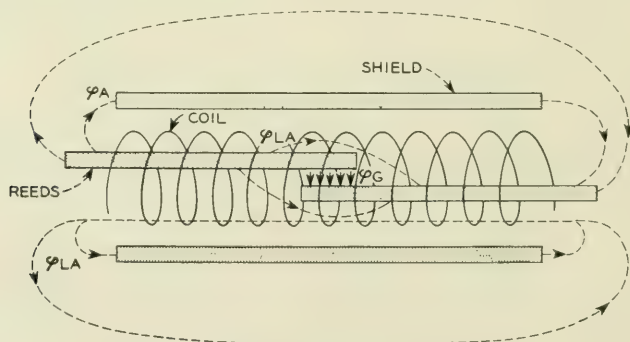


Fig. 24 — Magnetic field of a reed relay.

7 PULL EQUATIONS

Pull in Terms of Gap Flux

With the magnetization relations known, the pull can be determined from equations (7) or (8). Over the low density range, in which the magnetic circuit constants are independent of the flux, the field energy U is given by (5). Substitution of this in (7) gives the following expression for the pull F :

$$F = \frac{\varphi^2}{8\pi} \frac{d\mathfrak{R}}{dx}. \quad (26)$$

In terms of the equivalent circuit constants of Fig. 23, the reluctance $\mathfrak{R}$ is given by equation (24). Substituting this expression in (26) gives the equation:

$$F = \left(\frac{\mathfrak{R}_L}{\mathfrak{R}_0 + \mathfrak{R}_L + \frac{x}{A}} \right)^2 \frac{\varphi^2}{8\pi A}.$$

By comparison with (24), it can be seen that the bracketed term is the ratio $\mathfrak{R}_L/(\mathfrak{R}_0 + x/A)$, which equals the ratio of the gap flux to the total flux φ . Hence the preceding expression may be written in the form:

$$F = \frac{\varphi_g^2}{8\pi A} \quad (27)$$

By a parallel treatment it can be similarly shown that the application of equation (26) to the magnetic circuits of Fig. 3 and 21 gives expressions identical with (27) except that A is replaced by A_3 in the former case and by A_2 in the latter. Equation (27) is the familiar expression for

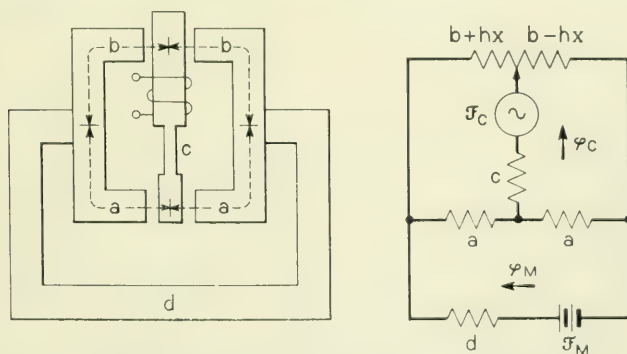


Fig. 25 — Magnetic circuit of a polar relay.

magnetic pull on two parallel planes with a field of uniform density φ_G between them. It can be directly derived from the fact that the gap energy is $\mathfrak{F}\varphi_G/(8\pi)$, with $\mathfrak{F} = x\varphi_G/A_2$. The change in this energy for a differential change dx of the gap equals $F dx$, so that F is given by equation (27). Derived in this way, it is apparent that (27) depends only on the field in the gap, and is quite independent of the density in the magnetic material.

Pull in Terms of Applied mmf

In the low density region, where $\mathfrak{R}$ is substantially a function of x only and the magnetization curves are linear, $W = U$, and hence the expression for F given by equation (8) becomes:

$$F = 2\pi(NI)^2 \frac{d}{dx} \left(\frac{1}{\mathfrak{R}} \right). \quad (28)$$

This expression for the pull in the linear region is known as the equation of Perrot and Picou.⁹ As $NI = \mathfrak{R}\varphi/(4\pi)$, it is identical with (26).

For the magnetic circuit of Fig. 23, substitution in (28) of the expression for $\mathfrak{R}$ given by (24) gives the following expression for the pull:

$$F = \frac{2\pi(NI)^2}{A \left(\mathfrak{R}_0 + \frac{x}{A} \right)^2}. \quad (29)$$

In the low density region, the reluctance can always be expressed in terms of the equivalent values of Fig. 25. Using the equivalent values of closed gap reluctance $\mathfrak{R}_0$ and of pole face area A , the pull is given by equation (29). This expression is therefore of general application in the low density region, and is the most convenient equation to use for this purpose.

High Density Pull

It was shown above that equation (27) can be derived directly from the expression for the field energy associated with the gap. This expression is quite independent of the reluctance in the rest of the magnetic circuit, and is therefore equally applicable in the low and high density regions. This essentially physical argument shows that (26) and (27) are applicable through the full range of magnetization.

The same result can be obtained from equation (7) by substituting in it the expression for U given by (3), and substituting $\mathfrak{R}\varphi/(4\pi)$ for NI .

There is thus obtained:

$$F = \frac{\partial}{\partial x} \int_0^\varphi \frac{\mathfrak{R}\varphi}{4\pi} d\varphi.$$

For saturation confined to the core, as in the series-parallel circuit of Fig. 21, $\mathfrak{R} = \mathfrak{R}_C + \mathfrak{R}_E$. On substituting this expression for $\mathfrak{R}$ in the preceding equation, $\mathfrak{R}_C$ is independent of x and $\mathfrak{R}_E$ is independent of φ , so that the expression becomes:

$$F = \frac{\varphi^2}{8\pi} \frac{d\mathfrak{R}_E}{dx},$$

or, for the series parallel circuit of Fig. 21

$$F = \frac{(\mathfrak{R}_{L2}\varphi)^2}{8\pi A_2 \left(\mathfrak{R}_{02} + \mathfrak{R}_{L2} + \frac{x}{A_2} \right)^2}. \quad (30)$$

Evidently this same expression should apply equally to the equivalent circuit of Fig. 23. That this is the case can be shown by substitution of the equivalence conditions of equations (25), giving an expression for the pull identical with (30) except that the magnetic circuit constants $\mathfrak{R}_{02}$, $\mathfrak{R}_{L2}$, and A_2 are replaced by their equivalent values $\mathfrak{R}_0$, $\mathfrak{R}_L$, and A .

Thus when saturation occurs in the core, as in most ordinary electro-magnets, (30) gives the pull through the full range of magnetization, and the expression is invariant to a change from the design constants of Fig. 21 to the equivalent constants of Fig. 23. If φ is taken as φ'' , the saturation flux, (30) gives the upper limit to the attainable pull. It is also useful, as shown in the companion article¹ on relay speed, in estimating the pull effective in rapid operation, when φ may be nearly constant during the latter stage of armature motion.

In estimating the steady state pull characteristics, equation (29) is applicable throughout the low density range. In the high density range, the pull may be determined from (27), or, provided saturation occurs in the core, from (30). In cases of armature saturation, the pull must be determined from (27), the expression of most general application.

8 SENSITIVITY AND WORK CAPACITY

Ampere Turn Sensitivity

Aside from any margin for rapid operation, the pull of the electro-magnet must exceed the static mechanical load at all values of gap x . As illustrated in Fig. 1, this may be studied graphically by comparing

the load curve with the family of pull curves for various values of NI . The just operate ampere turn value is that for the pull curve which just exceeds the load curve at all points. For design purposes, estimated pull and load characteristics must be used in this comparison, and conditions determined for minimizing the ampere turns required to operate a specified load.

The work V done against the load is represented by the area under the load curve. Its maximum possible value is equal to the work W done by the magnet, represented by the area under the pull curve. Within the region of substantially linear magnetization, the pull is given by (29), which may be written in the form:

$$F = \frac{F_0}{(1 + u)^2}, \quad (31)$$

where:

$$F_0 = \frac{2\pi(NI)^2}{A\mathcal{R}_0^2}, \quad (32)$$

and $u = x/(A\mathcal{R}_0)$, or the ratio of the gap reluctance x/A to the closed gap reluctance $\mathcal{R}_0$.

The work W done by the magnet in the travel from an initial gap x_1 to $x = 0$ is the integral of $F dx$, or of $A\mathcal{R}_0 F du$. From (31), this integral is given by:

$$W = \frac{u_1}{1 + u_1} W_{\max}, \quad (33)$$

where $u_1 = x_1/(A\mathcal{R}_0)$ and $W_{\max} = A\mathcal{R}_0 F_0$. For a large initial gap (u_1 large), W approaches $W_{\max}$, which therefore measures the upper limit to the output for this ampere turn value; the total area under the pull curve. In general, the difference between the force displacement characteristics of the load and pull curves permits only a fraction of $W_{\max}$ to be used to operate the load. The potential output of the magnet, however, depends solely upon $W_{\max}$. From (32), in which $A\mathcal{R}_0 F_0$ equals $W_{\max}$, the ampere turn sensitivity, or potential output for a given ampere turn value, is given by:

$$\frac{W_{\max}}{(NI)^2} = \frac{2\pi}{\mathcal{R}_0}. \quad (34)$$

Thus the ampere turn sensitivity depends solely upon the closed gap reluctance $\mathcal{R}_0$ throughout the region of linear magnetization.

Sensitivity for Specific Loads

The part of $W_{\max}$ that can be realized depends upon the shape of the load curve. It also depends upon the travel through which the load must be moved. In principle, the latter may be adjusted to any desired value by the choice of a lever arm determining the ratio of the travel at the point of load actuation to the travel at the point where x has been measured in determining the magnetic circuit constants. The use of a lever arm linkage, of course, changes the force required, but the work V and the shape of the load characteristic remain unchanged.

In Fig. 26 is shown the pull curve relation of equation (31), together with two simple load curves: a constant load and one varying linearly with travel. In the first case, let F_1 be the constant load required, and x_1 the required travel. The pull curve to operate this load must have $F = F_1$ at $x = x_1$. Hence, from (31), $F_1/F_0 = 1/(1 + u_1)^2$. The work V_c done against the load is $F_1 x_1$, and the ratio of V_c to $W_{\max}$ is therefore $F_1 x_1 (A\mathcal{R}_0 F_0)$, or $u_1 F_1 F_0$. Hence the part of $W_{\max}$ realized against a constant load is given by:

$$\frac{V_c}{W_{\max}} = \frac{u_1}{(1 + u_1)^2}. \quad (35)$$

The ratio $V_c/W_{\max}$ is shown plotted against u_1 in Fig. 27. Its maximum value is at $u_1 = 1$, where the travel $x_1 = A\mathcal{R}_0$. Maximum sensi-

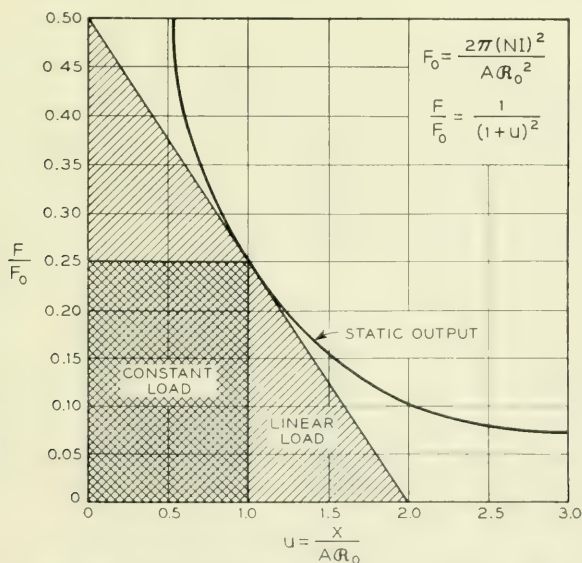


Fig. 26 — Relations between load and pull curves.

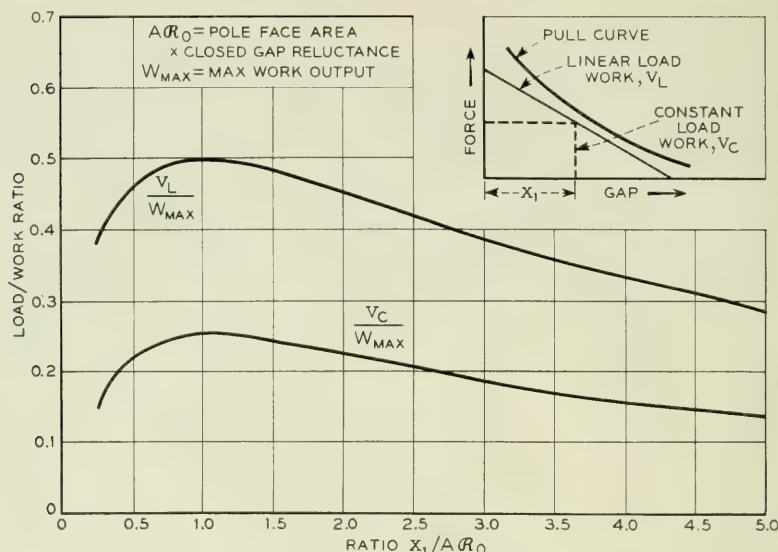


Fig. 27 — Relation between mechanical output and armature travel.

tivity is attained if the lever arm is chosen to satisfy this condition, and in this case $V_C = W_{max}/4$.

For the case of a linear load, varying from F_2 at $x = 0$ to zero at $x = x_2$, the load is given by:

$$F = F_2 \left(1 - \frac{u}{u_2} \right),$$

where $u_2 = x_2/(A R_0)$. For the pull curve to be tangent to this load curve, the values of both F and dF/dx given by this last equation and by (31) must be equal at the point of tangency, x_1 . These two conditions give two expressions for the ratio F_2/F_0 . Equating these, there is obtained the following equation for the point of tangency:

$$u_1 = \frac{2u_2 - 1}{3} \quad (36)$$

The work done against the load is $F_2 x_2/2$, and the ratio of this to W_{max} is therefore $F_2 x_2/(2A R_0 F_0)$. Substituting the expression for F_2/F_0 obtained as described, and the expression for u_1 given by (36), there is obtained the following expression for the fraction of W_{max} realized against a linear load:

$$\frac{V_L}{W_{max}} = \frac{(3u_1 + 1)^2}{4(u_1 + 1)^3}. \quad (37)$$

The ratio $V_L/W_{\max}$ is shown plotted against the point of tangency u_1 in Fig. 27. Its maximum value is at $u_1 = 1$, when $V_L = W_{\max}/2$. For this optimum condition, $u_2 = 2$, so that the total travel, x_2 , equals $2.4\theta_0$, while the point of tangency, x_1 , is at half that travel. For this linear load case, under the optimum condition, the maximum work that can be usefully applied is, then, $V_L = W_{\max}/2$.

Actual load curves seldom conform to either of the simple cases of Fig. 26, and more commonly have the irregular character illustrated in Fig. 1. Most of them, however, show a point of closest approach to the pull curve either at the junction of two segments, as in Fig. 28(a), or at a point of tangency to a linear segment, as in Fig. 28(b). In the former case, the coordinates of the junction point may be taken as F_1 and x_1 , as indicated, and the relations for a constant load for which $V_c = F_1x_1$ applied. In the other case, the tangent segment may be extended, as indicated, to intersect the axes at F_2 and x_2 , and the relations for a linear load for which $V_L = F_2x_2/2$ applied.

Quite generally, therefore, the work V done against the load curve is proportional to $W_{\max}$, and the proportionality constant is some function of $u_1 = x_1/(1.4\theta_0)$, where x_1 is the point of closest approach of the load and pull curves. Writing $f(u_1)/(2\pi)$ for the ratio $V/W_{\max}$, it follows from equation (34) that the ampere turn sensitivity, $V/(NI)^2$ is given by:

$$\frac{V}{(NI)^2} = \frac{f(u_1)}{\theta_0}, \quad (38)$$

where $f(u_1)$ is a maximum for $u_1 = 1$, and is similar to the curves of Fig. 27. For the particular load conditions represented by constant load

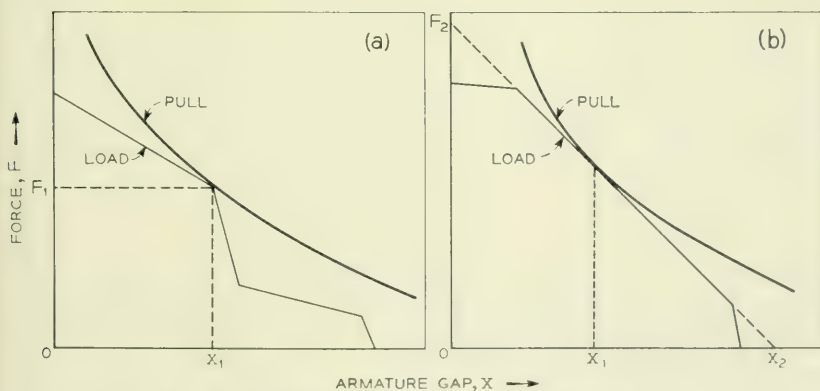


Fig. 28 Determination of pull curve to match load.

(V_c) and linear load (V_L) characteristics, the highest values of useful work were shown to be $\pi(NI)^2/(2\mathcal{R}_0)$, and $\pi(NI)^2/\mathcal{R}_0$, respectively.

These relations establish the important design requirement that the point of closest approach of the load and pull curves should be at $u_1 = 1$, or where the gap reluctance x_1/A equals the closed gap reluctance $\mathcal{R}_0$. To the extent permitted by space requirements and manufacturing considerations, this may be met by a proper choice of lever arm ratio or of pole face area. In the latter case $\mathcal{R}_0$ is not a wholly independent parameter but includes a term varying with A , so that allowance must be made for the change in $\mathcal{R}_0$ in changing A . The curves of Fig. 27 show that $V_c/W_{\max}$ and $V_L/W_{\max}$ decrease slowly for values of u_1 between 1 and 2. The ampere turn sensitivity is therefore close to its maximum value if u_1 lies in this range.

These relations are particularly useful in preliminary design estimates, as they permit the ampere turn requirement to be estimated for a known load merely from estimates of $\mathcal{R}_0$ and A . For this purpose it is only necessary to determine $V/W_{\max}$ by the procedure outlined above. With V known, this determines $W_{\max}$, which equals $A\mathcal{R}_0F_0$. Thus F_0 can be evaluated, and NI determined from equation (32).

Core Cross-Section

These relations are formally applicable only in the range of linear magnetization. For the estimates to be valid, however, it is only necessary that this condition be satisfied at the point of closest approach of load and pull curves, as an approach to saturation at smaller gaps will reduce the pull in a region where, in most cases, it is well in excess of the load curve.

For linear magnetization to obtain at the point of closest approach ($u = u_1$), the core flux should not materially exceed φ' , or aB' , where B' is the density for maximum permeability and a is the core cross-section. The flux is equal to $4\pi NI/\mathcal{R}(u_1)$, and the required value of a is therefore given by:

$$a = \frac{4\pi NI}{B'\mathcal{R}(u_1)}. \quad (39)$$

Here the values of NI and u_1 applying are those determined as described above, and $\mathcal{R}(u_1)$ is given by equation (24) for $x = u_1A\mathcal{R}_0$. To determine the core section a needed therefore requires an estimate of $\mathcal{R}_L$, as well as of $\mathcal{R}_0$ and A .

If the expression for NI given by (38) is substituted in (39), it is apparent that a varies as the square root of $V\mathcal{R}_0/f(u_1)$ and inversely as

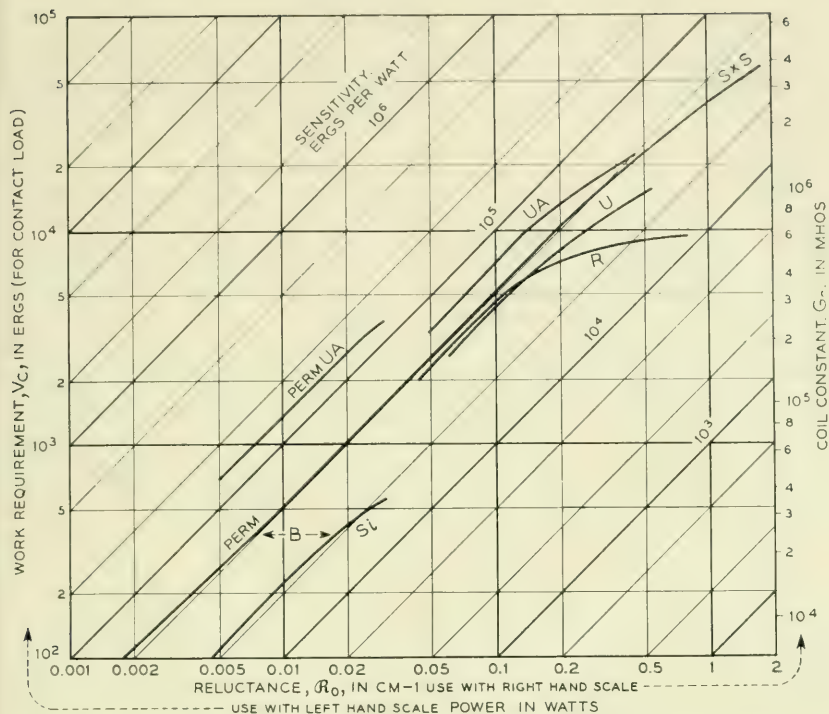


Fig. 29 — Work-power relations in terms of magnetic and coil constants.

$\mathfrak{R}(u_1)$. The latter is approximately proportional to $\mathfrak{R}_0$, so the core section a varies approximately as $1/\sqrt{\mathfrak{R}_0}$, and hence increases with increasing ampere turn sensitivity as well as with increasing load. For values of u_1 near unity, where $f(u_1)$ is nearly constant, a decreases as u_1 increases.

Power Sensitivity

The power sensitivity is the ratio of the work V done against the load to the steady state power I^2R . From equations (1) and (38), this is given by:

$$\begin{aligned} \frac{V}{I^2R} &= G_c f(u_1) / \mathfrak{R}_0, \\ &= \frac{eS}{\rho m^2} \frac{f(u_1)}{\mathfrak{R}_0}, \end{aligned} \quad (40)$$

where e/ρ is a constant determined by the coil construction, S is the coil volume, and m the mean length of turn. Depending on the coil depth,

the latter reflects the inside coil dimensions, which vary with the required core section a . The variation of the latter with u_1 discussed above has a minor effect on the power sensitivity, partially offsetting the effect of the term $f(u_1)$, so that $V/(NI)^2$ is nearly independent of u_1 in the range $1 \leq u_1 \leq 2$.

Ignoring these secondary variations, equation (40) shows that the steady state power required to operate a given load varies directly as the closed gap reluctance $\mathcal{R}_0$, and inversely as S/m^2 , where S is the coil volume and m the mean length of turn. Fig. 29 shows the relationship between the work that can be done, and the steady state power that must be supplied to the coil at this time. Since, in ordinary practice, use cannot be made of the entire area under the force-deflection curve, the "constant-load work," V_c is often used in comparing magnet designs. The maximum value for this term, as determined for a number of Western Electric relays, is shown in the figure. According to equation (40) the ratio of V_c to $I^2 R$ should here have the value

$$\frac{V_c}{I^2 R} = \frac{\pi G_c}{2 \mathcal{R}_0},$$

since for constant-load output $f(u_1) = \pi/2$, when the optimum work condition has been chosen.

Work Capacity

In applications where maximum sensitivity is not required, higher output may be obtained by operating the electromagnet through much of its travel in the non-linear region. The controlling factor is the flux developed at the point of closest approach of the load and pull curves, so the attainable output is controlled by the pull curve for constant flux, as given by equation (30). In what follows the flux will be taken as the saturation flux φ'' , corresponding to the upper limit of attainable output. The same relations may be used, however, to estimate the output attainable with any value of flux approaching φ'' in magnitude. For $\varphi = \varphi''$, (30) may be written in the form:

$$\frac{F}{F_0''} = \frac{C_L^2}{(C_L + u)^2}, \quad (41)$$

where $u = x/(A\mathcal{R}_0)$, $C_L = (\mathcal{R}_0 + \mathcal{R}_L)/\mathcal{R}_0$, and F_0'' is given by:

$$F_0'' = \frac{(C_L - 1)^2 \varphi''^2}{8\pi A C_L^2}. \quad (42)$$

Equation (41) is of the same form as equation (31), with u replaced

by u/C_L , and F_0 by F_0'' . Integration therefore gives an expression for the work W similar to equation (33), with u_1 replaced by u_1/C_L and $W_{\max}$, or $A\mathfrak{R}_0F_0$, replaced by $W_{\text{sat.}}$, or $C_L A\mathfrak{R}_0F_0''$. There is thus obtained:

$$W = \frac{u_1}{C_L + u_1} W_{\text{sat.}} \quad (43)$$

As (43) is of the same form as (33), with u replaced by u_1/C_L , the relations between the load and pull curves are the same in the two cases. Thus the ratios $V_C/W_{\text{sat.}}$ and $V_L/W_{\text{sat.}}$ may be obtained from (35) and (37), or from Fig. 27, by using u_1/C_L to replace u . Maximum output is thus obtained when $u_1/C_L = 1$, and the gap reluctance x_1/A at the point of closest approach equals C_L . Thus the optimum leverage for maximum output differs from that for maximum sensitivity by the factor C_L .

9 DISCUSSION

Applications

The applications of the magnetization relations considered in this article are confined to the static characteristics: the sensitivity and work capacity attainable when no specific timing requirements are imposed. As the same relations control both the electrical and mechanical response of an electromagnet under dynamic as well as under static conditions, the material outlined here has further applications in other aspects of magnet performance, as illustrated in the companion articles appearing in this issue of the JOURNAL.^{1,10}

The sensitivity and work capacity discussed above relate essentially to the operate characteristics. The other static characteristics of relay performance are those for release, and for marginal performance. The release characteristics are primarily dependent upon the operated load in relation to the pull at the closed gap. The rising pull characteristic as the gap is closed tends to give an operated pull well in excess of the operated load, giving a release ampere turn value small compared with the operate value. The coercive force tends to further decrease the release value, and imposes the need for stop pins to assure release. A high ratio of release to operate can be attained by providing a rising load characteristic to match the pull, by using high stop pins (thus increasing $\mathfrak{R}_0$), or by using a core or armature section that saturates in the latter part of the travel. A detailed discussion of the provisions for marginal operation is outside the scope of the present article, but obviously in-

volves application of the same relations that apply to the operate performance.

The procedures outlined for the estimation of sensitivity and work capacity, or of the magnetic parameters required to give a desired performance in these respects, have their primary use in preliminary design. They make it possible to determine from the performance requirements definite dimensional criteria which must be satisfied by the design. After models have been made and measured, the load characteristics may be compared directly with the observed pull. Study of the magnetization relations, and comparison of observed and estimated magnetic circuit constants, however, is of value through the whole course of development, and even in the engineering use of an established design. Such study relates the performance directly to the dimensions and materials used. In particular it permits extrapolation of the performance observed on particular models, providing estimates of the range of performance variation corresponding to the range of tolerances in dimensions and material properties. An essential phase of such studies is the experimental determination of the magnetic circuit constants by the methods discussed in a companion article.⁸

Validity

The solutions to the static field equations based on the magnetic circuit treatment are inherently approximations. The discussion of this subject in many engineering texts implies that the approximation is rather crude. In principle, as shown above, the magnetic circuit method may be applied to as close an approximation as desired, subject to the accuracy with which the pattern of the field can be recognized.

The criterion of satisfactory estimation of the magnetic circuit parameters is the ability to predict the magnetization relations actually observed. Analysis of observed magnetization relations by the method described in the companion article⁸ has shown that these relations can be satisfactorily represented by the magnetic circuit relations given here. Satisfactory agreement has been found between observed and estimated values of the magnetic circuit parameters when the latter are determined by the procedures described above. These conform in large measure to those initiated by Evershed,³ and are distinguished principally by explicitly recognizing the existence of leakage paths shunting any air gaps, and by recognizing that these cannot be identified by search coil measurements, so that the field between two members bounding a gap must be regarded as the vector sum of a constant and a variable field.

Design Considerations

The relations controlling the sensitivity and work capacity provide a direct indication of the effect of the configuration and dimensions of an electromagnet upon its performance. Here these will be considered only in general terms. Other things being equal, the sensitivity requirement determines the coil size, and hence the over-all dimensions, while the work capacity determines the cross-sections of the magnetic path. Within the framework of these generalizations, the dependence of performance on size and shape can be most readily considered in terms of the individual magnetic parameters.

The core reluctance $\mathcal{R}_c$ is characterized primarily by its minimum value $\mathcal{R}'_c$, and by the saturation flux φ'' which sets the upper limit to the work capacity. φ'' varies with the core cross-section a , as does φ' , the flux level for minimum reluctance. $\mathcal{R}'_c$ is given approximately by $\ell/(\mu'a)$, where ℓ is the core length, which varies with the coil size. For heavy duty relays with a large cross-section a , $\mathcal{R}_c$ is a minor component of the total reluctance except for φ very close to saturation. For highly sensitive light duty relays, ℓ is large and a small, making $\mathcal{R}_c$ large, and a must be made large enough for φ to be close to φ' , so that $\mathcal{R}_c$ may be close to its minimum value $\mathcal{R}'_c$.

The leakage reluctance $\mathcal{R}_L$ depends upon both the dimensions and the shape of the magnetic path. The leakage per unit length depends upon the ratio of the perimeter of the magnetic members to their separation, and is a minimum for a square magnet outline as in the two coil telegraph type relay. Space and accessibility considerations cause telephone relays to have a long narrow single coil configuration, increasing the leakage per unit length by a factor of from two to four over that of the square outline. Aside from this shape factor, the leakage reluctance varies inversely as the length. With respect to the static characteristics the leakage flux (1) increases the core cross-section required to attain a given work capacity, (2) decreases the equivalent pole face area, as shown by equations (25), (3) increases the equivalent closed gap reluctance $\mathcal{R}_0$, as shown by the same equations, thus decreasing the sensitivity. In addition, as discussed in the companion articles^{1,10} it delays operation and release by increasing the total field linked by the coil. In all these effects, the controlling factor is the ratio $\mathcal{R}_L/\mathcal{R}_0$, rather than the absolute value of $\mathcal{R}_L$, so these effects are most readily minimized by making $\mathcal{R}_0$ small.

The closed gap reluctance $\mathcal{R}_0$ is the sum of the reluctances of the return members and of the closed gaps and joints. The reluctance of the return members is comparable with that of the core but smaller, and in most

cases minor compared with that of the gaps and joints. The latter are equivalent to an air gap of 0.005 cm (2 mil-in) over the area of the joint, giving a marked advantage to one-piece construction of core and return members. The heel and main gaps necessarily introduce reluctances of similar magnitudes, further increased at the main gap by the height of the stop pins required to reduce the residual flux to the release level. It is therefore advantageous to use large areas for both heel and main gaps. When a small value of $\mathcal{R}_0$ is thus obtained, providing high sensitivity, its value is the more sensitive to variations in fit and alignment at the heel and main gaps, and high sensitivity therefore requires close tolerances on the dimensions controlling the fit of the armature to the pole pieces. Heavy section magnets tend to high sensitivity, partly because of the reduced reluctance of the magnetic members, but principally because heavy sections facilitate the provision of large areas at the gaps and joints.

As shown above, the effective pole face area is the reciprocal of the coefficient of x in the expression for the variable reluctance term. It therefore depends upon both the heel and main gaps, though the contribution from the heel gap is small when the armature is hinged there. For optimum sensitivity, or optimum work capacity, there are optimum values of the gap reluctance x_1/A as shown above, which can be obtained either by a choice of leverage to the load or by a choice of pole face area. It is preferable to attain these optima by varying the leverage, using as large a pole face area as possible, in order to make $\mathcal{R}_0$ small.

10 SUMMARY

The material given in this article provides a basis for relay design through the relationships between mechanical output and electrical input. From the indicated relations, one may find the mechanical work from a given magnet design, or find the magnetic design needed to provide a required mechanical output. Relationships for gaining optimum performance are also given.

The first step was to show that the mechanical work depends upon the field energy of the magnet, which is a consequence of the magnetomotive force provided. The magnetomotive force in turn is furnished by the electrical circuit, being determined jointly by the number of turns in the coil and the circuit resistance. The electrical output and the magnetic input were thus equated as

$$I^2 R = \frac{\mathfrak{F}^2}{16\pi^2 G_c},$$

where $G_c = N^2/R$, and is shown to be dependent on the coil dimensions and the conductivity of the wire in the winding.

The magnetic field energy is described by "magnetic flux" which, averaged for the entire magnet, is related to magnetomotive force through

$$\mathcal{F} = \varphi \mathfrak{R},$$

where $\mathfrak{R}$ is an expression accounting for the dimensions and materials of the magnet and its air gaps, called "magnetic reluctance." The validity of this magnetic circuit concept is discussed at some length, leading to the proof that for many cases a very simplified "equivalent two-mesh circuit" may be used to represent the more complicated actual cases. As a result, performance may be expressed in terms of the equivalent values: $\mathfrak{R}_0$, "closed gap reluctance"; $\mathfrak{R}_L$, "leakage reluctance"; and A , "pole face area." Values for these terms may be estimated for cases of initial design (magnet synthesis), or be measured so as to characterize completed models (magnet analysis). Evaluation of reluctances of various magnetic circuit components is described, with numerous relations given for gaps and for magnetic materials, including an approximate relation covering the non-linear behavior of ferrous materials.

In terms of the magnetic circuit variables, force and work may then be expressed as

$$F = \frac{\mathcal{F}^2}{8\pi A \mathfrak{R}_0^2} \frac{1}{(1+u)^2},$$

$$W = \frac{\mathcal{F}^2}{8\pi \mathfrak{R}_0} \frac{u}{1+u},$$

where $u = x/A\mathfrak{R}_0$, the ratio of air gap to the product $A\mathfrak{R}_0$. It is thus found that greatest magnet output for force systems common in relays is obtained when the critical load is picked up at a gap $x_0 = A\mathfrak{R}_0$, which may be accomplished by choice of lever arm. The maximum work that may thus be considered useful, $W_{\max}$, depends somewhat on the particular load characteristic, and for the typical constant-load case, V_c , is related to ampere turns, and to power input as follows:

$$V_c/(NI)^2 = \pi/2\mathfrak{R}_0,$$

$$V_c/I^2R = \pi G_c/2\mathfrak{R}_0.$$

Greatest useful magnet output is thus seen to depend directly on G_c , the "coil constant," and inversely on $\mathfrak{R}_0$, the equivalent closed gap reluctance.

Additional cases are given permitting design relations to be extended into the saturated range for the iron.

REFERENCES

1. Estimation and Control of Operate Time of Relays, Part I — Theory, R. L. Peek, Jr., page 109 of this issue. Part II — Applications, M. A. Logan, page 144 of this issue.
2. A. C. Keller, A New General Purpose Relay for Telephone Switching Systems, B.S.T.J., **31**, p. 1023, Nov. 1952.
3. S. Evershed, Permanent Magnets in Theory and Practice, Institute of Electrical Engineers Journal (England) **58**, 1919-1920.
4. H. C. Roters, *Electromagnetic Devices*, John Wiley and Sons, New York, 1941.
5. S. P. Thompson and K. W. Moss, Proc. Physical Society of London, **21**, p. 622, 1909.
6. W. B. Ellwood, Glass Enclosed Reed Relay, Elec. Eng. **66**, p. 1104, Nov., 1947.
7. E. B. Rosa and F. W. Grover, Bulletin Bureau of Standards, No. 169, 1912.
8. R. L. Peek, Jr., Analysis of Measured Magnetization and Pull Characteristics, page 79 of this issue.
9. Perrot and Picou, Comptes Rendus, **175**, Dec. 9, 1922.
10. R. L. Peek, Jr., Principles of Slow Release Relay Design, page 187 of this issue.

Analysis of Measured Magnetization and Pull Characteristics

By R. L. PEEK, Jr.

(Manuscript received September 21, 1953)

It is shown in this article that the observed magnetization relations of most ordinary electromagnets conform to simple expressions which can be interpreted as the flux vs. magnetomotive force equations of a reluctance network, analogous to the current-voltage equations of a resistance network. To the extent of such conformity the magnetic circuit constants characterizing the network suffice for the evaluation of the field energy and pull characteristics of the electromagnet. The agreement between the observed magnetization and these simple relations is close in the region of linear magnetization, and is adequate for engineering purposes at higher flux densities, but the extent of agreement in the latter range varies with the type of structure and the location of the magnetic parts which first approach saturation.

Specific analytical and graphical procedures are given for the evaluation of the magnetic circuit constants from both pull and magnetization measurements. These procedures employ relations which give linear plots indicating the degree of conformity of the observed relations to the expressions used to fit them. The relation of the measured constants to those which can be estimated in design¹ is discussed, as is the use and application of the measured constants in development and engineering studies.

1 INTRODUCTION

In the design of telephone relays and similar switching apparatus, the characteristics of the electromagnet which serves as motor element may be distinguished from those of the mechanical system of contact springs and actuating members which it operates. The performance of the electromagnet is characterized statically by the mechanical work done for a given coil energization, and dynamically by the time required to actuate the mechanical load, including in this the inertia of the moving parts.

Both the potential work output of the electromagnet and the energy stored in developing its field can be evaluated from its magnetization

relations. The consequent dependence of the pull and timing characteristics upon the magnetization relations makes the measurement and analysis of the latter fundamental to the understanding of performance and its relation to design.

Procedures are described in this article for the evaluation from measured magnetization relations of a few parameters which suffice for the determination of the pull and of the field energy under given conditions. These parameters, the magnetic circuit constants, characterize the electromagnet to which they apply. They are used as measures of performance in comparing different structures, or in studying the effect of dimensional or material variations in a particular design.

In addition to their significance as parameters summarizing measured magnetization relations, the magnetic circuit constants may be interpreted as the observed values of quantities postulated in the magnetic circuit approximation to static field theory. This approximation is discussed in a companion article,¹ which describes methods of estimating the values of the magnetic circuit constants from the configuration, dimensions, and materials constant of the electromagnet. Comparison of observed and estimated values of these constants has served as a guide in developing these methods of estimation. A complete design methodology is provided by the ability to both estimate and measure the magnetic circuit constants characterizing the performance of electromagnets.

In the experimental evaluation of the magnetic circuit constants described in this article, the basic method is that employing measured magnetization relations. Because of the dependency of the pull upon the magnetization relations, pull measurements may also be used, subject to certain limitations, for the evaluation of the magnetic circuit constants. The article includes description of procedures for doing this.

The notation used in this article conforms to the list that is given on page 257.

2 MAGNETIZATION RELATIONS

The magnetization relations of an electromagnet give the average flux ϕ linked per turn of the winding as a function of the two determining variables: applied ampere turns NI and armature position x . The relations are usually shown, as in Fig. 1, as a family of curves giving ϕ versus NI for various values of x . Armature motion is usually rotary, and the choice of the location at which x is measured is a matter of convenience. The curves of Fig. 1 apply to the electromagnet shown in Fig. 2, the AJ (heavy duty) type of wire spring relay described in an article

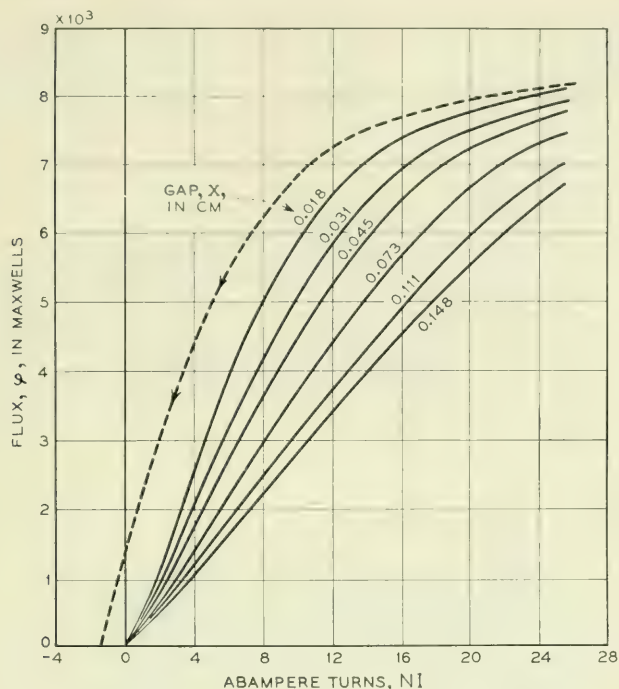


Fig. 1 — Magnetization curves.

by A. C. Keller.² In this case it is convenient to measure x at the center line of the actuating card, whose location with respect to the axis of rotation is indicated in Fig. 2.

Apart from interpretation in terms of magnetic theory, any observed point on a magnetization curve represents a measurement of the electrical energy U stored in the coil, as given by the integral

$$\int_0^I ei \, dt,$$

where $e = d(N\phi)/dt$. It is represented by the area between the magnetization curve and the axis of ϕ , and is given by:

$$U = \int_0^\phi NI \, d\phi. \quad (1)$$

This stored field energy may, in principle, be recovered in the decay of the field, except for the portion lost by hysteresis in the material, represented by the area between the increasing and decreasing mag-

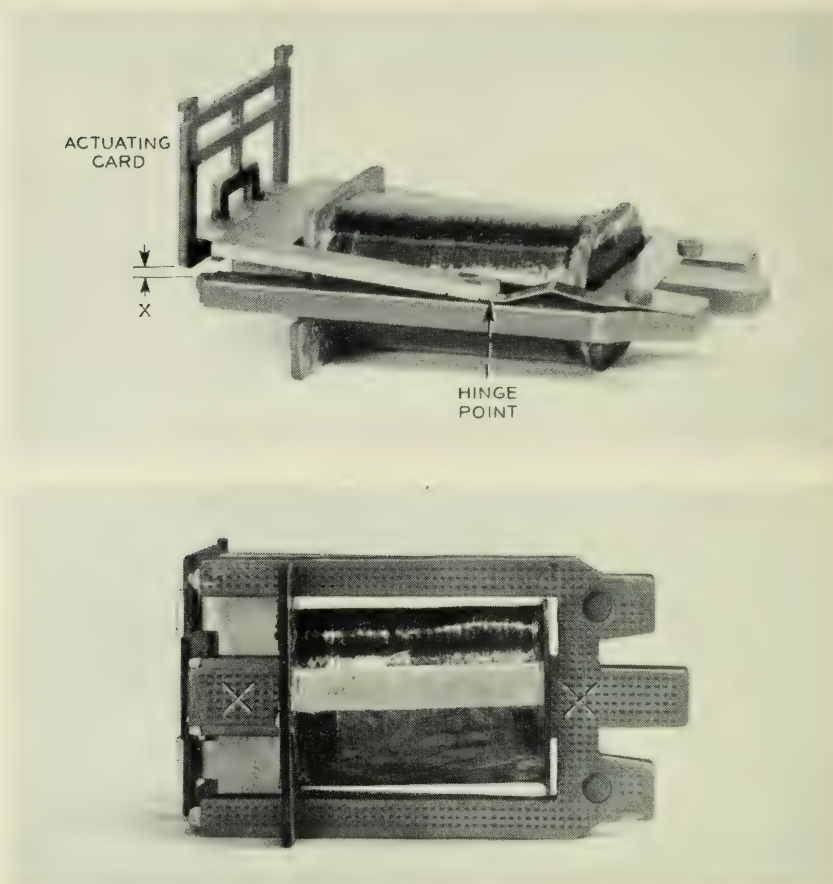


Fig. 2 — Magnetic parts, coil, and actuating card of AJ relay.

netization curves. The dashed curve of Fig. 1 is the decreasing magnetization curve for the same gap as the adjacent magnetization curve.

U is a function of x , and changes as the armature moves. If this motion occurs at a constant value of φ , there is no induced voltage and hence no electrical transfer of energy. A decrease in U must therefore represent mechanical work done on the armature. For a differential change dx in armature position, the mechanical work is Fdx , where F is the pull exerted on the armature. A general expression for F is therefore given by:

$$F = \frac{\partial U}{\partial x} \quad (\varphi \text{ constant}) \quad (2)$$

It follows from these considerations of energy balance that the magnetization relations determine the relations among the electrical input to the coil, the stored energy, and the mechanical output attainable. Hence, any generally applicable and simple expression for the magnetization relations, even if purely empirical, would provide a convenient means for evaluating and describing the performance characteristics.

The expressions given below for the magnetization relations are not purely empirical, but are those obtained as approximate solutions to the magnetic field equations by the magnetic circuit method. When experimentally evaluated, however, their utility in defining the characteristics of the electromagnet to which they apply is independent of this interpretation, and depends only on their conformity to the observed magnetization relations.

It is convenient to formulate the expressions for the magnetization curves in terms of the reluctance $\mathcal{R}$, the ratio $\mathcal{F}/\phi$, where $\mathcal{F}$ is the magnetomotive force $4\pi NI$. The observations of Fig. 1 are plotted in Fig. 3 in the form of curves giving $\mathcal{R}$ vs. NI for various values of x . A constant value of ϕ is represented in such a plot by a straight line through the origin. The radial lines used as supplementary co-ordinates are spaced to give a convenient scale for ϕ .

The reluctance curves of Fig. 3 are similar in character to those applying to most ordinary electromagnets. Each curve has a relatively flat characteristic in the vicinity of a minimum located on a common flux

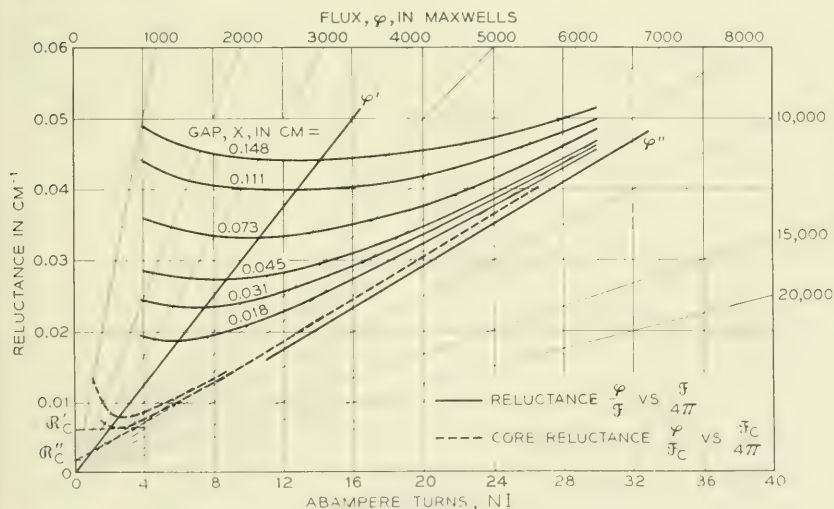


Fig. 3 — Reluctance curves.

line, that marked φ' in the figure. At values of φ above this minimum, the reluctance increases at an increasing rate, indicating an upper limit to the value of φ approached as NI becomes very large. The observed value of φ' may be interpreted as that for which the core density corresponds to maximum permeability, while the indicated upper limit may be interpreted as the saturation flux φ'' .

The observations plotted in Figs. 1 and 3 were obtained with the sample initially demagnetized, a convenient reference condition for measurement. In actual use, relays and other electromagnets have been previously operated, and the applicable φ versus $\mathcal{F}$ relation is that for repeated magnetization. In this case there is little or no increase in reluctance below φ' . For engineering purposes, the reluctance below φ' may be taken as constant at the value observed at φ' in measurements made from an initially demagnetized condition. This assumption is equivalent to taking the φ versus $\mathcal{F}$ relation as linear up to the "knee" of the curve, the point of tangency with a line through the origin.

It follows that expressions for the magnetization relations may take $\mathcal{R}$ as a function of x only, independent of φ , in the low density region, $\varphi < \varphi'$, while for $\varphi > \varphi'$, modified expressions must be employed which show $\mathcal{R}$ increasing with φ , and approaching $\mathcal{F}/\varphi''$.

Magnetic Circuit Schematics

The reluctance expressions for the magnetization relations can be conveniently defined by the magnetic circuit schematics shown in Figs. 4, 5, and 6. In these the reluctance $\mathcal{R}$, or $\mathcal{F}/\varphi$, is represented by a network of component reluctances, analogous to a network of electrical resistances, with flux analogous to current, and magnetic potential analogous to voltage. Thus the reluctance corresponding to Fig. 4 is given by:

$$\mathcal{R} = \frac{\mathcal{R}_L \left(\mathcal{R}_0 + \frac{x}{A} \right)}{\mathcal{R}_L + \mathcal{R}_0 + \frac{x}{A}}. \quad (3)$$

The interpretation of the parameters $\mathcal{R}_0$, $\mathcal{R}_L$, and A in terms of the magnetic circuit concept is discussed in the companion article cited above.¹ In the analysis of experimental results they are parameters which summarize measurements in which the observed values of $\mathcal{R}$ conform to (3). From this viewpoint Fig. 4 indicates only that the total flux φ is the sum of a flux φ_L for which the reluctance is constant, and a flux φ_G for which the reluctance varies directly with x .

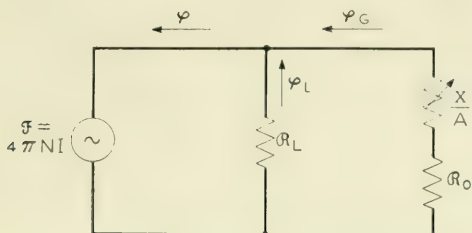


Fig. 4 — Equivalent magnetic circuit.

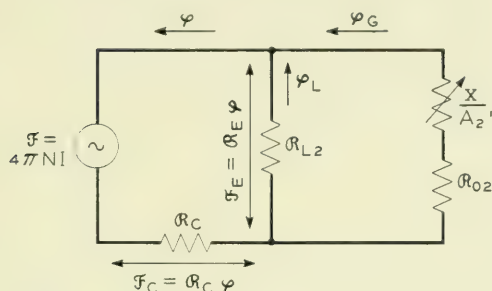


Fig. 5 — Magnetic circuit for core saturation.

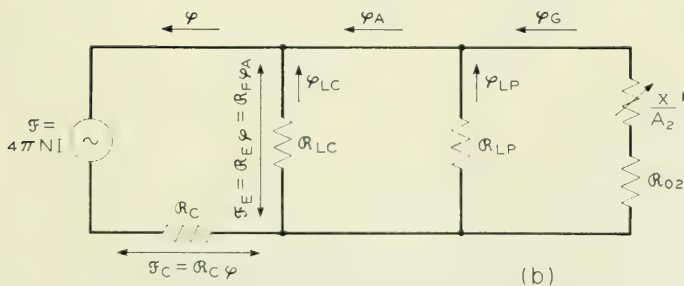
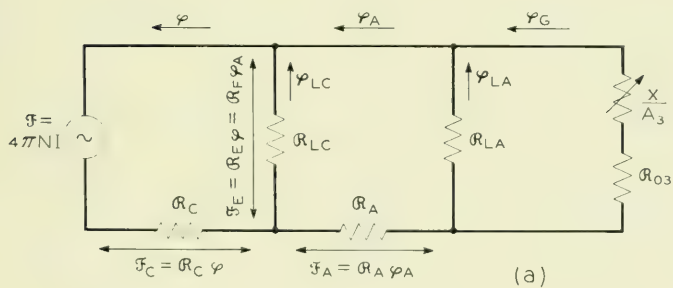


Fig. 6 — Magnetic circuit for armature saturation.

In Fig. 4, $\mathcal{R}_0$, $\mathcal{R}_L$, and A are constants, so that $\mathcal{R}$ is a function of x only, independent of φ . Equation (3) can then apply only to the low density region, $\varphi < \varphi'$. For most ordinary electromagnets, the magnetization relations for $\varphi < \varphi'$ conform to (3), and hence to the schematic of Fig. 4.

For $\varphi > \varphi'$, the expression for $\mathcal{R}$ must provide for the variation with φ . It usually suffices to use an expression in which the only term varying with φ is that corresponding to the path in which saturation first occurs. The most common case is that in which saturation first occurs in the core. The reluctance can then be taken as conforming to the schematic of Fig. 5, in which $\mathcal{R}_{02}$, $\mathcal{R}_{L2}$ and A_2 are constants, and $\mathcal{R}_C$ is a function of φ only. The total reluctance is given by:

$$\mathcal{R} = \mathcal{R}_C + \mathcal{R}_E, \quad (4)$$

where:

$$\mathcal{R}_E = \frac{\mathcal{R}_{L2} \left(\mathcal{R}_{02} + \frac{x}{A_2} \right)}{\mathcal{R}_{L2} + \mathcal{R}_{02} + \frac{x}{A_2}}. \quad (5)$$

If the variation with φ is taken as conforming to the empirical Frohlich-Kennelly equation,⁶ $\mathcal{R}_C$ is given by:

$$\mathcal{R}_C = \mathcal{R}_C'' \frac{\varphi''}{\varphi'' - \varphi} = \mathcal{R}_C' \frac{\varphi'' - \varphi'}{\varphi'' - \varphi}, \quad (6)$$

where φ'' is the saturation flux and φ' is the flux for maximum permeability or minimum reluctance $\mathcal{R}_C'$. If these assumptions apply, the minimum values of $\mathcal{R}$ for all values of x must lie on φ' , as for the results of Fig. 3. This common minimum is thus evidence that the core is controlling with respect to the variation with φ , and that (4) and (6) are applicable. As (6) applies only for $\varphi > \varphi'$, $\mathcal{R}_C'$ is the value of $\mathcal{R}_C$ not only at $\varphi = \varphi'$, but throughout the low density region $\varphi < \varphi'$. In the alternative form given by (6), $\mathcal{R}_C''$ is defined by this equation, and represents merely the intercept of a plot of $\mathcal{R}_C$ extended below the region to which (6) applies.

In some electromagnets saturation occurs in the armature rather than in the core. This is the case, for example, in high speed relays in which the armature section is minimized to reduce its mass. In such cases, the reluctance conforms to the schematic of Fig. 6(a), in which $\mathcal{R}_A$ represents the reluctance of the armature, which varies with φ_A . As the ratio φ_A/φ

then varies with x , the minimum values of $\mathfrak{R}$ do not lie on a common value of φ in this case.

In Fig. 6(a), $\mathfrak{R}_A$ may be taken as given by an expression of the same form as (6), with φ' and φ'' replaced by the minimum and saturation values of φ_A , and $\mathfrak{R}'_C$ replaced by the minimum value of $\mathfrak{R}_A$. $\mathfrak{R}_C$ may be taken either as given by (6), or as a constant if the variation in $\mathfrak{R}_A$ is dominant. The other parameters of Fig. 6(a) are constants. The expression for $\mathfrak{R}$ applying to Fig. 6(a) may be obtained in the same manner as those given for the circuits of Figs. 4 and 5.

The magnetic circuits of Figs. 5 and 6 are called "design" circuits, because their constants may be estimated from the dimensions and material constants of the design, as discussed in the companion article.¹ Estimates of the corresponding values of the equivalent magnetic circuit constants of Fig. 4 are obtained from these design constant estimates by the equivalence equations given below.

Conditions of Equivalence

Whatever magnetic circuit is taken as applying, the component reluctances are independent of φ in the low density region, $\varphi < \varphi'$, and are therefore constants except for the gap reluctance. If this varies directly as x , as assumed in the schematics shown, the reluctance for any circuit reduces to an expression of the form of (3), applying to the circuit of Fig. 4. As shown in the companion article,¹ the reluctance of the circuit of Fig. 5 when $\mathfrak{R}_C$ is constant is given by (3) when the parameters of this equation are given by:

$$\begin{aligned}\mathfrak{R}_L &= \mathfrak{R}_{L2} + \mathfrak{R}_C, \\ A &= \frac{A_2}{p^2}, \\ \mathfrak{R}_0 &= p^2 \mathfrak{R}_{02} + p \mathfrak{R}_C,\end{aligned}\tag{7}$$

where:

$$p = 1 + \frac{\mathfrak{R}_C}{\mathfrak{R}_{L2}}.$$

These relations are the conditions of equivalence, for which the reluctances of Figs. 4 and 5 are identical for all values of x . With them, the reluctance given by (4) can be reduced to the simpler form of (3) when $\mathfrak{R}_C$ is a constant, as throughout the low density region.

Similar relations apply to the magnetic circuit of Fig. 6(a). In particu-

lar, when $\mathcal{R}_A$ is a constant, the reluctance $\mathcal{R}_F$, or $\mathcal{F}_E/\varphi_A$, represented by the series parallel path of φ_A in Fig. 6(a), may be represented by the simple parallel path of φ_A in Fig. 6(b). By analogy with equations (7) the constants applying are related by:

$$\begin{aligned}\mathcal{R}_{LP} &= \mathcal{R}_{LA} + \mathcal{R}_A, \\ A_2 &= \frac{A_3}{p^2}, \\ \mathcal{R}_{02} &= p^2 \mathcal{R}_{03} + p \mathcal{R}_A,\end{aligned}\tag{7A}$$

where

$$p = 1 + \frac{\mathcal{R}_A}{\mathcal{R}_{LA}}.$$

The circuit of Fig. 6(b) is identical with that of Fig. 5 with $1/\mathcal{R}_{L2}$ equal to the sum of $1/\mathcal{R}_{LC}$ and $1/\mathcal{R}_{LP}$. Thus the circuit of Fig. 6(a) may be reduced to that of Fig. 5 when $\mathcal{R}_A$ is constant. The resulting circuit may be reduced in turn to the equivalent circuit of Fig. 4 when $\mathcal{R}_C$ is constant. Thus the equivalent circuit may be used to represent all cases in the low density region, where the component reluctances are independent of φ , provided the gap reluctance varies linearly with x .

3 ANALYSIS OF MAGNETIZATION MEASUREMENTS

In analysing the observed magnetization relations, it is convenient to treat the low density and high density relations in separate and successive steps. For the former, the analysis is based on the equivalent magnetic circuit of Fig. 4.

EVALUATION OF EQUIVALENT MAGNETIC CIRCUIT CONSTANTS

For a given value of x , the total reluctance in the low density region is taken as constant at the minimum value $\mathcal{R}'(x)$ observed in measurements made from the demagnetized condition. If the observations are plotted as reluctance curves, as in Fig. 3, these values of $\mathcal{R}'(x)$ may be read directly. When the recording fluxmeter³ is used, $\mathcal{R}'(x)$ can be obtained from the φ versus $\mathcal{F}$ curve as the slope $\mathcal{F}/\varphi$ of the line through the origin tangent to the curve. The reciprocals $P(x)$ of the values of $\mathcal{R}'(x)$ thus evaluated may be plotted against x . Values of $P(x)$ corresponding to the minimum reluctance values of Fig. 3 are plotted in this way in Fig. 7. The origin of x is taken as for a gap of 0.025 cm., corresponding to the operated position of the actuating card for maximum stop pin height.

Values of x measured from this origin are termed travel, as distinguished from gap values, which are measured from the position of iron to iron contact.

If the relation between $\alpha'(x)$ and x conforms to (3), the values of $P(x)$ must conform to:

$$P(x) = \frac{1}{\alpha_L} + \frac{A}{A\alpha_0 + x}. \quad (8)$$

Then if x_c is a central reference value of x , and $P(x_c)$ the corresponding value of $P(x)$, the expression for $P(x_c) - P(x)$ given by (8) may be written in the form:

$$\frac{x - x_c}{P(x_c) - P(x)} = A \left(\alpha_0 + \frac{x_c}{A} \right)^2 + \left(\alpha_0 + \frac{x_c}{A} \right) (x - x_c). \quad (9)$$

Thus conformity with (3) requires the ratio given by (9) to vary linearly with x . This ratio is plotted against $x - x_c$ in Fig. 7, referred to the upper and right hand scales. The values of this ratio are sensitive to deviations in the values of x and $P(x)$ near x_c , and minor variations from linearity are to be expected here. Allowing for this, the results agree with (9), and the relation between $\alpha'(x)$ and x therefore conforms to (3).

From (9), the slope and intercept of the linear plot may be interpreted as indicated in Fig. 7. Then A may be evaluated by dividing the intercept

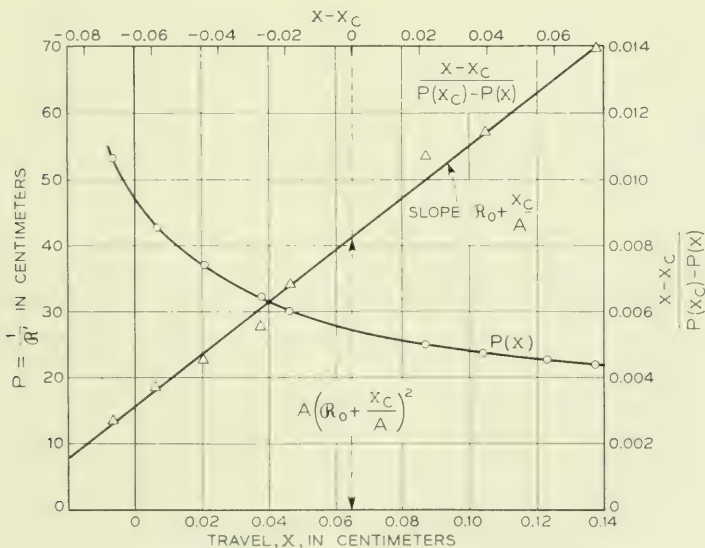


Fig. 7 — Evaluation of equivalent circuit constants.

value by the square of the slope value, and the latter then serves to evaluate $\mathcal{R}_0$ by subtracting x_c/A . On substituting these values of $\mathcal{R}_0$ and A in (8) at $x = x_c$, $\mathcal{R}_L$ may be evaluated. For the case plotted in Fig. 7, the values thus obtained are:

$$\mathcal{R}_0 : 0.0300 \text{ cm}^{-1}$$

$$A : 1.34 \text{ cm}^2$$

$$\mathcal{R}_L : 0.0690 \text{ cm}^{-1}$$

These values of the equivalent circuit constants, substituted in (3), suffice to accurately characterize the magnetization relations and thus the electromagnet's performance through the low density region.

Evaluation of High Density Relations

For cases of core saturation, the high density relations may be analysed in terms of the magnetic circuit of Fig. 5. The corresponding reluctance is given by (4), in which $\mathcal{R}_C = \mathcal{R}'_C$ for $\varphi = \varphi'$. If $\mathcal{R}'_C$ can be determined, the three parameters determining $\mathcal{R}_E$ in (5) may be evaluated by the equivalence equations (7). To determine $\mathcal{R}_C$ through the high density region from (6) requires evaluation of $\mathcal{R}'_C$, φ' , and φ'' . Thus the analysis of the high density relations requires evaluation of these three quantities in addition to the three parameters of the equivalent circuit evaluated above.

A method for evaluating $\mathcal{R}'_C$, φ' , and φ'' directly from the magnetization curves is described subsequently, following a description of a preferred method which requires additional measurements. These measurements are determinations of the drop $\mathcal{R}_{c\varphi}$ in magnetic potential through the core, and are conveniently made with the magnetomotive force gauge developed by W. B. Ellwood.⁴

Magnetomotive Force Measurements

The Ellwood mmf gauge measures the difference in magnetic potential between the outer ends of the two reeds of a dry reed switch, as determined by the coil current required to open the switch. The end of one reed is used as a probe, which is brought into contact with the magnetic structure being measured. The end of the other reed, at a distance of about 5 cm from the probe, then lies on a relatively distant surface surrounding the structure: this surface is substantially at the same potential for any position of the gauge. Then the difference between two such measurements made at different points on the magnetic structure measures the difference in magnetic potential between them.

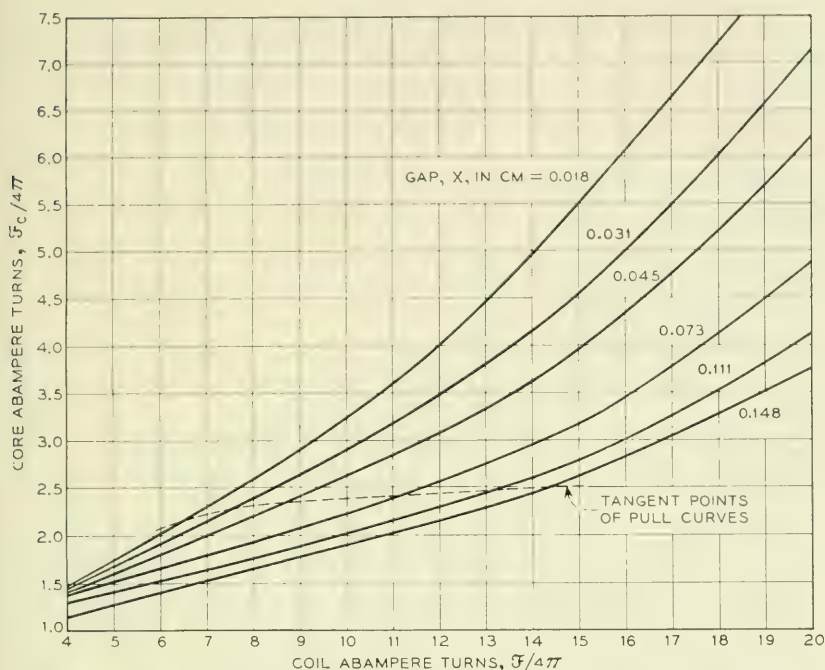


Fig. 8 — Relation of core potential drop to applied magnetomotive force.

To the extent that the physical structure may be identified with the corresponding magnetic circuit element, the drop in magnetic potential in the core may be identified with $\mathcal{R}_c\varphi$ in Fig. 5. Assuming this identification to apply, magnetic probe measurements may be made at the two ends of the core, as at the points marked X in Fig. 2. The potential drop observed in these measurements is the magnetic potential $\mathfrak{F}_E$ applied to the external magnetic circuit. $\mathfrak{F}_E$ is equal to the applied magnetomotive force $\mathfrak{F}$ (or $4\pi NI$) less the potential drop $\mathfrak{F}_c = \mathcal{R}_c\varphi$ in the core. Thus $\mathfrak{F}_c = \mathfrak{F} - \mathfrak{F}_E$. Values of $\mathfrak{F}_c/(4\pi)$ thus determined for the relay of Fig. 2 are shown plotted against $\mathfrak{F}/(4\pi)$ for various values of x in Fig. 8. These curves are substantially linear in the low density region; their upward concavity at higher values of $\mathfrak{F}$ is evidence that saturation first occurs in the core.

Evaluation of Core Reluctance Constants

The magnetization curves of Fig. 1 and the core magnetomotive force curves of Fig. 8 were obtained with the same model. For given values of $\mathfrak{F}$ and x there can be read from these two figures corresponding values of

φ and $\mathcal{F}_c$, giving the corresponding values of core reluctance $\mathcal{R}_c = \mathcal{F}_c/\varphi$. Such values of $\mathcal{F}_c$ have been determined for two values of x (the smallest and largest included in these figures), and are shown plotted against $\mathcal{F}_c/(4\pi)$ in the dashed curves included in Fig. 3. Values of $\mathcal{R}_c$ for intermediate values of x are intermediate between these two curves, whose similarity supports the assumption of Fig. 5 and equation (4) that $\mathcal{R}_c$ is substantially independent of x .

For $\varphi > \varphi'$ the plot of $\mathcal{R}_c$ versus $\mathcal{F}_c$ is substantially linear. In this it conforms to (6) which, by substitution of $\mathcal{F}_c/\mathcal{R}_c$ for φ , may be written in the form:

$$\mathcal{R}_c = \mathcal{R}_c'' + \frac{\mathcal{F}_c}{\varphi''} = \left(1 - \frac{\varphi'}{\varphi''}\right) \mathcal{R}_c' + \frac{\mathcal{F}_c}{\varphi''}. \quad (10)$$

Thus φ'' may be evaluated from the slope of the plot of $\mathcal{R}_c$ versus $\mathcal{F}_c$, and $\mathcal{R}_c''$ from the intercept of the dotted line extension, as indicated in Fig. 3. The observed relation deviates from (10) in the vicinity of φ' , and the linear relation conforming to (10) does not intersect $\mathcal{R}_c'$ at φ' , but at some higher value of φ . Thus (6) more nearly represents the observed relation if φ' is taken as this higher value, rather than at the flux for minimum $\mathcal{R}$ previously taken as φ' and marked at such in Fig. 3. The error in using this latter value of φ' in (6), however, is minor, and of little significance in engineering estimates.

Alternative Method of Determining Core Reluctance Constants

When magnetomotive force measurements are not made, the core reluctance constants can be determined directly from the magnetization curves as follows:

As can be seen in Fig. 3, the line representing φ'' can be directly identified as the apparent asymptote of the reluctance curves: specifically by determining the flux line parallel to the tangent to the upper portion of the lowest reluctance curve. As φ' is the value of φ at which the reluctance curves have their common minima, φ' and φ'' can be readily evaluated, and only $\mathcal{R}_c'$ remains to be determined.

On substituting $\mathcal{F}/\mathcal{R}$ for φ in (6), and substituting this expression for $\mathcal{R}_c$ in (4) the resulting equation can be solved to give:

$$\mathcal{R} = \frac{1}{2} \left(\mathcal{R}_E + \mathcal{R}_c'' + \frac{\mathcal{F}}{\varphi''} + \sqrt{\left(\mathcal{R}_E + \mathcal{R}_c'' - \frac{\mathcal{F}}{\varphi''} \right)^2 + 4\mathcal{R}_c'' \frac{\mathcal{F}}{\varphi''}} \right).$$

Let $\mathcal{R}''$ be the particular value of $\mathcal{R}$ for which

$$\mathcal{F} = \mathcal{R}_c' \varphi'' = (\mathcal{R}_c' + \mathcal{R}_E) \varphi''.$$

As both $\mathcal{R}'_c/\mathcal{R}'$ and $\mathcal{R}''_c/\mathcal{R}'$ are small, substitution of $\mathcal{F} = \mathcal{R}'\varphi''$ in the preceding expression gives as an approximation:

$$\mathcal{R}'' = \mathcal{R}' + \sqrt{\mathcal{R}''_c \mathcal{R}'},$$

from which:

$$\mathcal{R}''_c = \frac{(\mathcal{R}'' - \mathcal{R}')^2}{\mathcal{R}'}.$$
(11)

Thus, as indicated in Fig. 9, the value of $\mathcal{R}'$ may be read from any one of the reluctance curves, and the corresponding value of $\varphi''\mathcal{R}'$ computed. The value of $\mathcal{R}$ corresponding to $\mathcal{F} = \mathcal{R}'\varphi''$ is read from the curve, and taken as $\mathcal{R}''$. The values of $\mathcal{R}''$ and $\mathcal{R}'$ are then substituted in (11) to determine $\mathcal{R}''_c$. This procedure may be followed for all the reluctance curves, and a mean value of $\mathcal{R}''_c$ computed. $\mathcal{R}'_c$ is then taken as

$$\varphi''\mathcal{R}''_c/(\varphi'' - \varphi').$$

Evaluation of External Reluctance Constants

The procedures described above provide for the determination (a) of the equivalent constants of Fig. 4, (b) of the minimum value $\mathcal{R}'_c$ of $\mathcal{R}_c$ in Fig. 5. The remaining constants of Fig. 5 can then be computed by means of the equivalence equations (7). Table I lists, for the measurements of Figs. 1 and 8, the values of the equivalent constants deter-

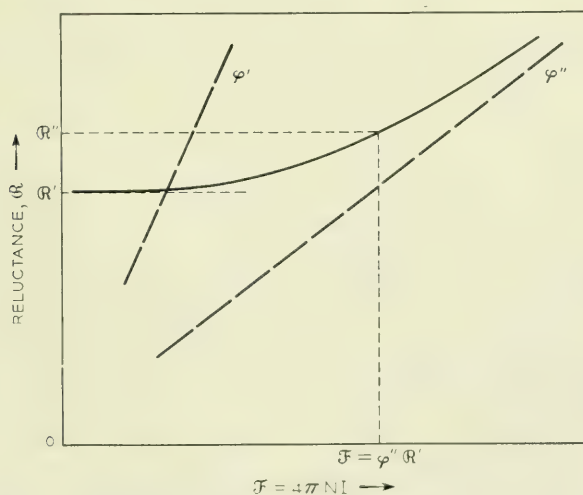


Fig. 9 — Alternative method of evaluating core reluctance.

TABLE I — EVALUATION OF MAGNETIC CIRCUIT CONSTANTS

Equivalent Values (Fig. 4)	Design Values (Fig. 5)
From Fig. 7	From Fig. 3
$\mathcal{R}_0 : 0.0300 \text{ cm}^{-1}$	$\mathcal{R}'_C : 0.0070 \text{ cm}^{-1}$
$A : 1.34 \text{ cm}^2$	From equations (7)
$\mathcal{R}_L : 0.0690 \text{ cm}^{-1}$	$\mathcal{R}_{02} : 0.0181 \text{ cm}^{-1}$
	$A_2 : 1.66 \text{ cm}^2$
	$\mathcal{R}_{L2} : 0.0620 \text{ cm}^{-1}$

mined in Fig. 7, and the value of $\mathcal{R}'_C$ determined in Fig. 3. These values have been substituted in equations (7) to determine the component terms of $\mathcal{R}_E$.

The constants applying to Fig. 5 are of interest primarily for comparison with the values computed from the design, as discussed in the companion article.¹ While formally required to evaluate the high density magnetization relations, they are not explicitly required in estimating the high density pull, as is shown in a later section.

The reluctance $\mathcal{R}_E$ of equations (4) and (5) can be evaluated experimentally through the full range of the magnetization measurements when the magnetomotive force measurements are available. Thus Figs. 1 and 3 can be used to determine values of φ and $\mathcal{F}_E$ (or $\mathcal{F} - \mathcal{F}_C$) for corresponding values of $\mathcal{F}$ and x . In this way there have been obtained the curves of $\mathcal{R}_E$ plotted against $\mathcal{F}_E$ for various values of x in Fig. 10. These curves are included here merely to illustrate the approximate validity of (4), in which $\mathcal{R}_E$ is taken as independent of φ . The departures from this condition are minor except for the smaller gaps at high values

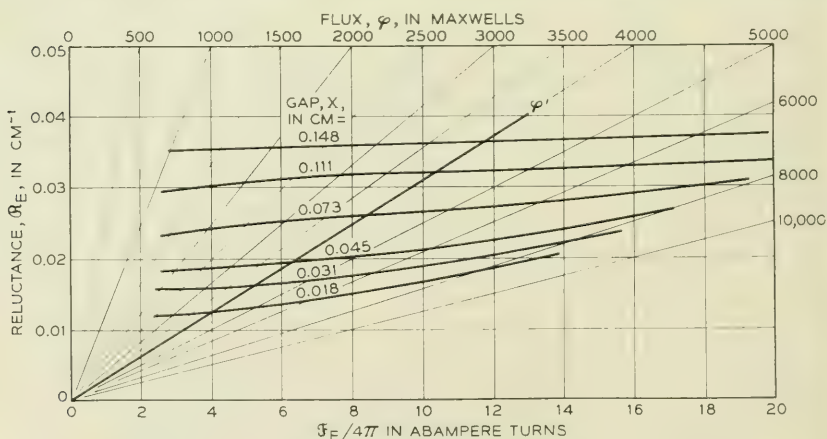


Fig. 10 — Reluctance curves for magnetic path external to the core.

of φ . Here the increase in flux density in the iron parts external to the core results in a significant increase in $\mathcal{R}_E$ as φ increases. By comparison with the plot of $\mathcal{R}_C$ included in Fig. 1, however, this increase in $\mathcal{R}_E$ is minor, and does not affect the fact that the limiting reluctance is $\mathfrak{F}_E \varphi''$, where φ'' is the saturation flux of the core.

It may be noted in passing that the values of $\mathcal{R}_E$ lying on φ' must conform to (5). As this is of the same form as (3), these values of $\mathcal{R}_E$ may be used to evaluate the component terms of (5) by the procedure applied to the values of $\mathcal{R}'$ in Fig. 7. As the same data are employed, the values of $\mathcal{R}_{02}$, $\mathcal{R}_{L2}$, and A_2 thus determined agree with those computed by means of equations (7) within the accuracy of the computations.

Case of Armature Saturation

The analysis of the low density relations is, of course, independent of where saturation occurs, and involves merely the determination of the equivalent magnetic circuit constants by the procedure described above. In the high density region, however, the quantities appearing in Fig. 6(a) must be evaluated when armature saturation occurs. This requires measurements not only of $\mathfrak{F}$ and φ , but also of $\mathfrak{F}_E$ and φ_A .

$\mathfrak{F}_E$ is given by the Ellwood mmf gauge measurements previously described. To measure φ_A requires the use of a search coil wound on the armature, located over the region of maximum density. For twin return path structures, such as the relay of Fig. 2, twin search coils must be used, connected to measure the total armature flux. If the variation of $\mathcal{R}_A$ with φ_A is to be determined directly, additional mmf gauge measurements must be made of the potential drop $\mathcal{R}_A \varphi_A$ through the armature.

With φ_A determined as a function of $\mathfrak{F}_E$ for various values of x , the reluctance $\mathcal{R}_F$, or $\mathfrak{F}_E \varphi_A$ may be analysed by the procedure described above for the analysis of the reluctance $\mathcal{R}$ or $\mathfrak{F} \varphi$. The curves of $\mathcal{R}_F$ versus $\mathfrak{F}_E$ have minima $\mathcal{R}'_F$ at φ'_A . The reciprocals of these values of $\mathcal{R}'_F$ are plotted and analysed as in Fig. 7 to evaluate the equivalent constants $\mathcal{R}_{LP}$, $\mathcal{R}_{02}$, and A_2 of Fig. 6(b). The constants $\mathcal{R}'_A$, φ'_A , and φ''_A characterizing the relation between $\mathcal{R}_A$ and φ_A are determined either from mmf gauge measurements or by the method of Fig. 9. Then the constants $\mathcal{R}_{LA}$, $\mathcal{R}_{03}$, and A_3 of Fig. 6(a) can be evaluated by means of the equivalence equations (7A).

To complete the determination of the quantities appearing in Fig. 6(a), $\mathcal{R}_{LC}$ is evaluated as the ratio $\mathfrak{F}_E (\varphi - \varphi_A)$, while $\mathcal{R}_C$, which may either be constant or conform to (6), is evaluated from the relation between φ and $\mathfrak{F}_C$.

4. PULL RELATIONS

As shown in Section 2, the pull F can be determined from the field energy U by means of equation (2). If the gap reluctance varies linearly with x , this equation reduces to:

$$F = \frac{\varphi_g^2}{8\pi A}, \quad (12)$$

as shown in the companion article.¹ Equation (12) is Maxwell's law for the pull between parallel plane surfaces of area A . As used here, $1/A$ has the more general sense of the coefficient of x in a linear expression for the gap reluctance. This equation is therefore applicable to all the cases previously discussed, with A , of course, taken as A_2 or A_3 for the relations applying to Figs. 5 and 6. As with the magnetization relations, it is convenient to give separate consideration to the pull at low densities and that at high densities.

Pull for Linear Magnetization

At low densities, where the magnetization relations are approximately linear, the reluctance conforms to the equivalent magnetic circuit of Fig. 4. In this case, φ_g is given by $\mathfrak{F}/(\mathfrak{R}_0 + x/A)$, and equation (12) becomes:

$$F = \frac{\mathfrak{F}^2}{8\pi A \left(\mathfrak{R}_0 + \frac{x}{A} \right)^2},$$

which may be written in the more familiar form:

$$F = \frac{2\pi(NI)^2}{A \left(\mathfrak{R}_0 + \frac{x}{A} \right)^2}. \quad (13)$$

For conformity with (13) the pull for a given value of x should vary as $(NI)^2$, giving a linear plot with a slope of two when plotted against NI on logarithmic paper. Fig. 11 shows pull measurements of the relay of Fig. 2 plotted in this way. The dashed lines tangent to the curves have a slope of two, and conform in this respect to (13). It is convenient to denote the pull indicated by these dashed lines as F' . Near and below the point of tangency the actual pull F differs little from F' . The pull, like the magnetization, is measured under the condition of initial demagnetization, which gives lower magnetization, and hence less pull, than normally applies in actual use when the magnet has been previously

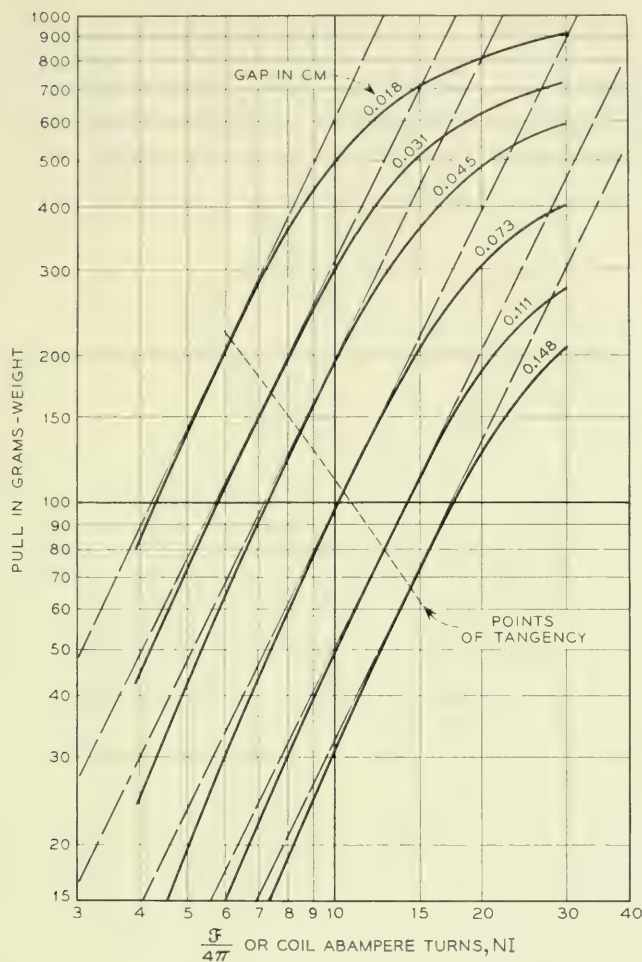


Fig. 11 — Pull curves.

operated. Hence F' satisfactorily represents the pull applying in practice through the low density region. In the high density region, the pull falls off as the result of incipient saturation, and the ratio F'/F increases as $\mathfrak{F}$ (or $4\pi NI$) is increased.

High Density Pull — Core Saturation

It is convenient to express the high density pull in terms of the ratio F'/F , where F is the actual pull, and F' the indicated pull conforming to (13) for the same value of $\mathfrak{F}$ (or $4\pi NI$). As F'/F is unity at low densi-

ties, it follows from (12) that F'/F is equal to $\bar{\varphi}_G^2/\varphi_G^2$, where φ_G is the actual value, and $\bar{\varphi}_G$ the fictitious value that would obtain if the magnetic circuit reluctances remained constant at all values of $\mathfrak{F}$. For the case of core saturation, to which the magnetic circuit of Fig. 5 applies, the ratio φ_G/φ is constant for a given value of x , and hence in this case F'/F equals $\bar{\varphi}^2/\varphi^2$ where φ is the actual value and $\bar{\varphi}$ the fictitious value that would obtain if the reluctance were $\mathfrak{R}'$. Thus for core saturation,

$$\frac{F'}{F} = \left(\frac{1}{\varphi} \frac{\mathfrak{F}}{\mathfrak{R}'} \right)^2,$$

and hence:

$$\frac{F'}{F} = \left(\frac{\mathfrak{R}}{\mathfrak{R}'} \right)^2. \quad (14)$$

As values of $\mathfrak{R}'$ and $\mathfrak{R}$ can be read directly from magnetization curves, this relation provides a simple means of computing the high density pull for specific values of x and $\mathfrak{F}$, provided the low density pull has been measured or estimated from (13), using the values of $\mathfrak{R}_0$ and A determined from the low density magnetization relations.

An approximate expression for the right-hand side of (14) can be obtained in a convenient form from the two limiting conditions: $\mathfrak{R} = \mathfrak{R}'$ for $\mathfrak{F} = \mathfrak{R}'\varphi'$, and $\mathfrak{R}$ approaches $\mathfrak{F}/\varphi''$ as $\mathfrak{F}$ becomes very large. A simple expression satisfying these conditions is:

$$\mathfrak{R}^2 = \mathfrak{R}'^2 \left(1 - \left(\frac{\varphi'}{\varphi''} \right)^2 \right) + \frac{\mathfrak{F}^2}{\varphi''^2}.$$

Aside from meeting the limiting conditions, this, like the Frohlich-Kennelly relation used previously, is a purely empirical expression. Assuming it to apply, the high density pull in the case of core saturation is, from (14), given by:

$$\frac{F'}{F} = 1 - \left(\frac{\varphi'}{\varphi''} \right)^2 + \left(\frac{\mathfrak{F}}{\mathfrak{R}'\varphi''} \right)^2. \quad (15)$$

As F' is given by (13), this expression gives the high density pull in terms of the magnetomotive force $\mathfrak{F}$, or explicitly:

$$F = \frac{\mathfrak{F}^2}{8\pi A \left(\mathfrak{R}_0 + \frac{x}{A} \right)^2 \left(1 - \left(\frac{\varphi'}{\varphi''} \right)^2 + \left(\frac{\mathfrak{F}}{\mathfrak{R}'\varphi''} \right)^2 \right)}, \quad (15A)$$

where $\mathfrak{R}'$ is given by (3), so that the constant terms in this expression are confined to φ' , φ'' , and the equivalent magnetic circuit constants A , $\mathfrak{R}_0$, and $\mathfrak{R}_L$.

High Density Pull — Armature Saturation

A similar treatment is applicable in the case of armature saturation, where F'/F is equal to φ_G^2/φ_A^2 as before. In this case, however, it is φ_G/φ_A rather than φ_G/φ that is constant, so that the expression corresponding to (14) is:

$$\frac{F'}{F} = \left(\frac{\Re_F}{\Re_F'} \right)^2, \quad (16)$$

where $\Re_F$ is the reluctance $\mathfrak{F}_E/\varphi_A$ of Fig. 6(a), and $\Re_F'$ is its minimum value. The limiting expression for $\Re_F$ is $\mathfrak{F}_E/\varphi_A''$, and the approximation corresponding to (15) is therefore:

$$\frac{F'}{F} = 1 - \left(\frac{\varphi_A'}{\varphi_A''} \right)^2 + \left(\frac{\mathfrak{F}_E}{\Re_F' \varphi_A''} \right)^2. \quad (17)$$

In general, there is no simple expression for $\mathfrak{F}_E/\mathfrak{F}$, and (17) is of little interest as a general expression. When saturation is wholly confined to the armature, however, $\Re_c$ is a constant and usually a minor term, and $\mathfrak{F}_E$ is then nearly equal to $\mathfrak{F}$. In this case, the high density pull is given by the approximation:

$$F = \frac{\mathfrak{F}^2}{8\pi A \left(\Re_0 + \frac{x}{A} \right)^2 \left[1 - \left(\frac{\varphi_A'}{\varphi_A''} \right)^2 + \left(\frac{\mathfrak{F}^2}{\Re_F' \varphi_A''} \right)^2 \right]}. \quad (17A)$$

The application of this approximation may be simplified by developing an approximate expression for $\Re_F'$ in terms of the equivalent magnetic circuit constants. The expression for $\Re_F'$ can be read from Fig. 6(b), where $\Re_{LP}$ is usually large compared with $\Re_{02} + x/A_2$, so that the latter term is an approximate expression for $\Re_F'$. As (17A) applies only for $\Re_c$ small, and as Fig. 6(b) is then of nearly the same form as the equivalent circuit of Fig. 4, the equivalent constants $\Re_0$ and A are in this case approximately equal to $\Re_{02}$ and A_2 , respectively. Hence in the case of saturation confined wholly to the armature, to which (17A) applies, R_F' can be taken as given by $\Re_0 + x/A$, and (17A) then involves only φ_A' , φ_A'' , and the equivalent magnetic circuit constants $\Re_0$ and A .

5 ANALYSIS OF PULL MEASUREMENTS

The following discussion relates primarily to the use of pull measurements in analytical studies as a means of evaluating the magnetic circuit constants.

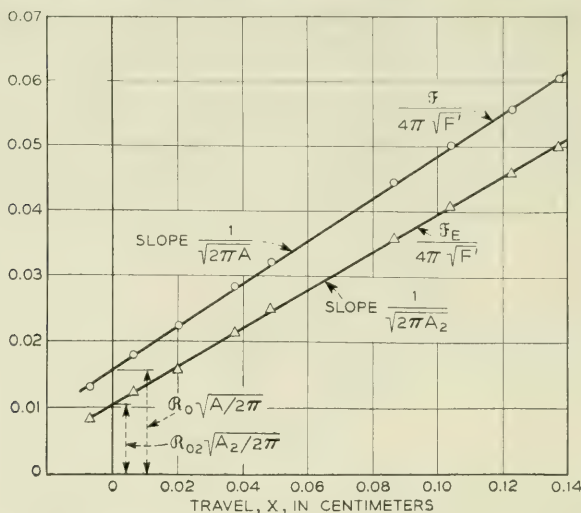


Fig. 12 — Evaluation of magnetic circuit constants from pull data.

Evaluation of Equivalent Magnetic Circuit Constants

The low density pull is given approximately by equation (13), to which the dashed F' lines of Fig. 11 conform. These lines coincide with the actual pull curves at the points of tangency, which must correspond to the minimum reluctance values. Taking F as F' , (13) may be written in the form:

$$\frac{\mathfrak{F}}{4\pi\sqrt{F'}} = \mathcal{R}_0 \sqrt{\frac{A}{2\pi}} + \frac{x}{\sqrt{2\pi A}}. \quad (18)$$

Values of NI/F' have been determined for the dashed lines of Fig. 11, and are plotted against x in Fig. 12 — the plot marked $\mathfrak{F}/(4\pi\sqrt{F'})$. In agreement with (18) these points determine a straight line. The slope and intercept have been used to determine the values of $\mathcal{R}_0$ and A listed under "Pull Results" in Table II.

As the reluctance $\mathcal{R}_E$ in the circuit of Fig. 5 is identical in form with the reluctance $\mathcal{R}$ of Fig. 4, the pull for Fig. 5 is given by an expression similar to (13), with $\mathfrak{F}$, $\mathcal{R}_0$, and A replaced by $\mathfrak{F}_E$, $\mathcal{R}_{02}$, and A_2 , respectively. As the two circuits are equivalent at the points of minimum reluctance and the corresponding points of tangency for the pull curves, the pull at these points is the same for the two cases. Thus if the ratio of $\mathfrak{F}_E$ to $\mathfrak{F}$ is determined at the points of tangency, a plot of $\mathfrak{F}_E/(4\pi\sqrt{F'})$ against x should conform to:

$$\frac{\mathfrak{F}_E}{4\pi\sqrt{F'}} = \mathcal{R}_{02} \sqrt{\frac{A_2}{2\pi}} + \frac{x}{\sqrt{2\pi A_2}}. \quad (19)$$

TABLE II — EVALUATION OF MAGNETIC CIRCUIT CONSTANTS

From Pull Results	From Magnetization Results
$\mathfrak{R}_0 : 0.0319 \text{ cm}^{-1}$	0.0300 cm^{-1}
$A : 1.46 \text{ cm}^2$	1.34 cm^2
$\mathfrak{R}_{02} : 0.0188 \text{ cm}^{-1}$	0.0181 cm^{-1}
$A_2 : 1.85 \text{ cm}^2$	1.66 cm^2
$\mathfrak{R}_L : 0.0630 \text{ cm}^{-1}$	0.0690 cm^{-1}
$\mathfrak{R}_{L2} : 0.0560 \text{ cm}^{-1}$	0.0620 cm^{-1}
$\mathfrak{R}_C : 0.0072 \text{ cm}^{-1}$	0.0070 cm^{-1}

The points of tangency are indicated on the pull curves of Fig. 11. These values of $\mathfrak{F}$ have been marked on the corresponding curves of Fig. 8, from which can be determined the corresponding values of $\mathfrak{F}_C$, and thus of $\mathfrak{F}_E$. It may be noted in passing that these values of $\mathfrak{F}_C$ are all similar, corresponding to a mean level of about 24 ampere turns, or 30 gilberts. The magnetization results gave values of 4,000 maxwells for φ' (Fig. 3) and of 0.007 cm^{-1} for $\mathfrak{R}_C$ (Table I), giving a value of 28 gilberts for the product. This agrees with the value of $\mathfrak{F}_C$ at the tangent points of the pull curves, showing that these points coincide with the points of minimum reluctance ($\varphi = \varphi'$).

Taking the ratio of $\mathfrak{F}_E$ to $\mathfrak{F}$ as read from the curves of Fig. 8 at the indicated tangent points, the values of $\mathfrak{F}'/(4\pi\sqrt{F'})$ in Fig. 12 have been multiplied by the corresponding values of $\mathfrak{F}_E/\mathfrak{F}$, giving the points indicated by triangles in this figure. These conform to a straight line relation, in agreement with (19). The slope and intercept of this linear plot have been used to determine the values of $\mathfrak{R}_{02}$ and A_2 listed under "Pull Results" in Table II.

The four quantities determined from the two plots of Fig. 12 appear in the three independent equivalence equations (7), and these three equations may therefore be used to determine the other three quantities: $\mathfrak{R}_L$, $\mathfrak{R}_{L2}$, and $\mathfrak{R}_C'$. The resulting values of these latter quantities are listed under "Pull Results" in Table II. For comparison, the table includes the corresponding values of these quantities obtained from the magnetization measurements, and given previously in Table I. The agreement between the results derived from these different kinds of measurements indicates the validity of the analysis described here.

High Density Pull

As previously noted, it is convenient to express the pull F in the high density region in terms of the ratio F'/F , where F' is the pull computed from (13) for the value of $\mathfrak{F}$ applying. F'/F is readily evaluated from a logarithmic plot of the observed pull versus $\mathfrak{F}$, as illustrated by Fig. 11.

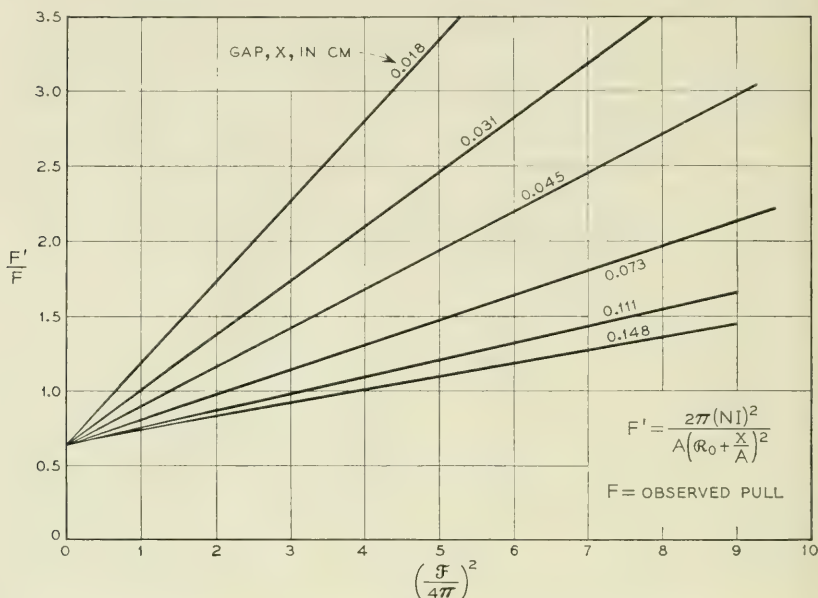


Fig. 13 - Pull characteristics for approaching saturation.

The dashed tangent lines give values of F' , so corresponding values of F' and F can be read for various values of $\mathfrak{F}$. Values of F'/F for the results of Fig. 11 have been thus determined, and are plotted in Fig. 13 against $(\mathfrak{F}/4\pi)^2$ for the values of x applying to the individual curves of Fig. 11.

The values of F'/F thus plotted determine a family of straight lines having a common intercept at the origin of $\mathfrak{F}^2$. In this respect these results are representative of the pull characteristics of electromagnets when plotted in this way. As an empirical fact, apart from any analysis, this linearity provides a useful method of plotting pull observations in the high density region, as it facilitates interpolation and the detection of inconsistent observations.

The observed linearity in the relation between F'/F and $\mathfrak{F}^2$ conforms to the relations postulated in (15) and (17), of which the former should apply for core saturation. If (15) applies, the common intercept of 0.65 at $\mathfrak{F}^2 = 0$ in Fig. 13 should equal $1 - (\phi'/\phi'')^2$, indicating a value for ϕ'/ϕ'' of 0.59. To determine if the relation fully conforms to (15), the relation between the slopes of the lines and the corresponding values of x must be compared with that indicated by this equation.

Let y^2 be the observed slope of the plot of F'/F versus $\mathfrak{F}^2$ for a particu-

lar value of x . If the relation conforms to (15), $y = 1 - (\alpha' \varphi'')$, where α' is given by (3). Writing C_L for the ratio $(\alpha_L + \alpha_0) / \alpha_0$, and u for the ratio $x / (A\alpha_0)$,

$$\alpha' = \frac{1 + u}{C_L + u} \alpha_L,$$

and hence the values of y should conform to the equation:

$$y(1 + u) = \frac{C_L + u}{\alpha_L \varphi''}. \quad (20)$$

Using the values of A and α_0 determined from the pull results in Fig. 12 and tabulated in Table II, values of u have been determined for the values of x applying in Fig. 13, and the corresponding values of y have been determined from the slopes of the corresponding lines. The values of $y(1 + u)$ thus determined are shown plotted against u in Fig. 14. The

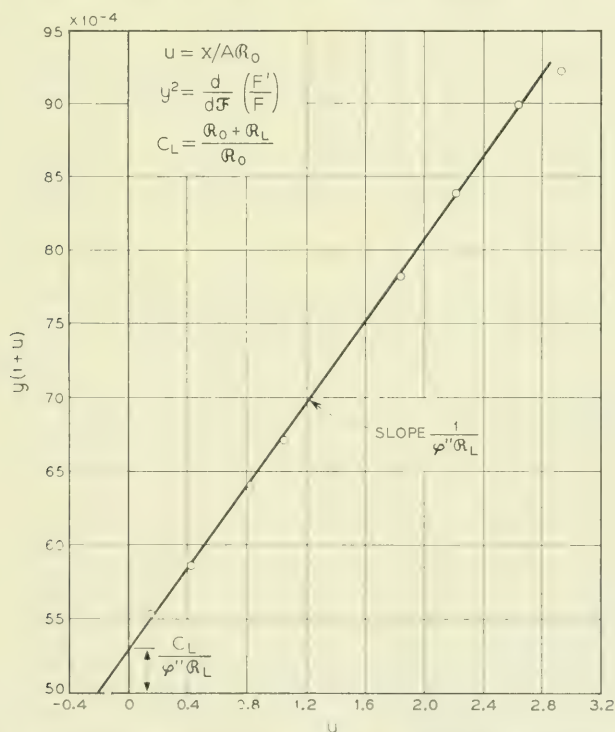


Fig. 14 — Evaluation of saturation flux φ'' and leakage reluctance α_L from pull characteristics.

TABLE III — EVALUATION OF MAGNETIC CIRCUIT CONSTANTS

From Pull Measurements	From Magnetization Measurements
Figs. 13 & 14: φ' : 4,700 maxwells φ'' : 8,000 maxwells $\mathcal{R}_L$: 0.091 cm^{-1}	Fig. 3: φ' : 4,000 maxwells φ'' : 8,500 maxwells
Fig. 12: A : 1.46 cm^2 $\mathcal{R}_0$: 0.0319 cm^{-1} $\mathcal{R}_L$: 0.063 cm^{-1}	Fig. 7: A : 1.34 cm^2 $\mathcal{R}_0$: 0.0300 cm^{-1} $\mathcal{R}_L$: 0.069 cm^{-1}

resulting plot is linear, in agreement with (20), showing the observed relation between F'/F and $\mathfrak{F}^2$ to conform fully to (15).

The expressions for the slope and intercept given by (20) are indicated in Fig. 14. From the values of the slope and intercept given by the plot, C_L may be evaluated, giving $\mathcal{R}_L$, as $\mathcal{R}_0$ is known. With $\mathcal{R}_L$ thus evaluated, φ'' may be determined from the value of the slope. The values of $\mathcal{R}_L$ and φ'' thus determined from results of Fig. 14 are listed in Table III, together with the value of φ' given by the ratio φ' / φ'' determined from the common intercept in Fig. 13. Included in this table under "Pull Results" are the values of A , $\mathcal{R}_0$, and $\mathcal{R}_L$ determined from the low density data and previously tabulated in Table II. Values of the same quantities as determined from the magnetization measurements are included for comparison.

For both pull and magnetization analysis, empirical relations have been assumed in (6) and (15) for the variation of $\mathcal{R}$ with φ , agreeing with each other and the known conditions only at their upper and lower limits. In view of this, the agreement in the values of φ' and φ'' between the pull and magnetization results is good, and the deviation in the value of $\mathcal{R}_L$ found in Fig. 14 from the other values of this quantity is not surprising.

High Density Pull for Armature Saturation

The low density pull for armature saturation may be analysed as in the case of core saturation. The pull data may be plotted against $\mathfrak{F}$ on logarithmic paper as in Fig. 11, and the tangents of slope two giving F' drawn, as in this figure. The values of $\mathfrak{F}$, $\sqrt{F'}$ may be plotted as in Fig. 12 to determine A and $\mathcal{R}_0$, and A_2 and $\mathcal{R}_{02}$ also evaluated if $\mathfrak{F}_E$ has been determined from magnetomotive force gauge measurements. As before, the remaining magnetic circuit constants of Figs. 4 and 5 may be de-

terminated: the latter in this case are those corresponding to the equivalent schematic of Fig. 6(b).

For the high density pull, values of F'/F may be determined and plotted against $\bar{x}$ as before. The resulting plot is similar to Fig. 13 in showing linear relations for each value of x , radiating from a common intercept. For the presentation of pull results, therefore, this method of plotting is applicable for both core and armature saturation, and for the case where saturation is approached concurrently in both.

When the slopes y'' are determined, however, and $y(1 + u)$ plotted against u as in Fig. 14, a straight line is not usually obtained unless saturation is confined to the core. The observed relation for armature saturation is usually a curve which is concave upward, and approaches the horizontal at small values of u (travel x small). This result can be interpreted from the expression (17) for F'/F in the case of saturation confined to the armature. The corresponding expression for y is $1/(\mathcal{R}'_F \varphi''_A)$. It was shown above that in this case $\mathcal{R}'_F$ is approximately equal to $\mathcal{R}_0 + x/A$ or $\mathcal{R}_0(1 + u)$. The corresponding expression for $y(1 + u)$ is therefore the constant term $1/(\mathcal{R}_0 \varphi''_A)$. Thus for saturation confined to the armature, the plot of $y(1 + u)$ versus u is simply a horizontal line, from whose value φ''_A can be determined. In the more usual case of concurrent saturation, this relation is approached at small values of u , where φ_A/φ is relatively large, while at larger values of u , where φ_A/φ is smaller, the relation approaches that for core saturation.

6 DISCUSSION

The procedures for the analysis of pull and magnetization measurements described in this article are primarily intended for the evaluation of the magnetic circuit constants in development studies. They also have some related applications, and the choice of procedure varies with the application and with the measurements available. The following discussion of the applications of this analysis indicates the most convenient procedures for each case.

Presentation of Pull Measurements

Pull measurements are used not only for the guidance of development studies, but as engineering data in the application of relays and other electromagnets. A convenient form of presentation is that used in Fig. 11, where the pull is plotted against ampere turns on logarithmic paper. In preparing such data, interpolation of the measurements and the

recognition of inconsistent observations is facilitated by a supplementary plot of F'/F versus $\mathfrak{F}^2$ (or $(NI)^2$) as in Fig. 13.

Estimation of Pull

Estimates of pull are made in preliminary development either from estimates of the magnetic circuit constants, computed from the dimensions as discussed in the companion article,¹ or from magnetization measurements. The low density pull is given by (13), requiring only the equivalent circuit constants A and $\mathfrak{R}_0$ for its evaluation, and is represented by the F' lines of slope two in a logarithmic plot of F versus NI . Estimates of the high density pull may be prepared as linear plots of F'/F versus $(NI)^2$. As these conform (for core saturation) to (15) they require estimates of ϕ' , ϕ'' , and of $\mathfrak{R}'$, as given by (3).

Estimates of pull may also be required in preparing pull data for engineering use, with respect to the changes in pull associated with dimensional and other variations in a commercial product. Such estimates may be obtained by estimating the effect of the variations on the magnetic circuit constants fitting the observed pull, and computing the pull for the changed constants in the manner described in the preceding paragraph.

Evaluation of Equivalent Magnetic Circuit Constants

The equivalent magnetic circuit constants, which define the field energy and the pull characteristics through the low density range, largely suffice for the prediction of performance. This is illustrated by the use of these constants in the analysis of capability relations in one companion article¹ and in the dynamic relations presented in another.⁵

These constants are most readily and accurately evaluated from magnetization data by the procedure illustrated in Fig. 7 and the associated computations. A check on the values of A and $\mathfrak{R}_0$ is provided by analysing pull data by the procedure illustrated in Fig. 12. When only pull data are available, A and $\mathfrak{R}_0$ can be determined in this way, but the determination of $\mathfrak{R}_L$ requires the measurement of $\mathfrak{F}_E$ by the Ellwood mmf gauge. With $\mathfrak{F}_E$ known, the pull results suffice for the determination not only of $\mathfrak{R}_L$, but of the design circuit constants of Fig. 5 (provided saturation is confined to the core).

Evaluation of the Magnetic Design Circuit Constants

The magnetic design circuit constants are required only for comparison with estimated values, in development studies of the effect on the

performance of the configuration, dimensions and materials of the electromagnet. They are most readily and accurately determined from coil magnetization measurements. It is convenient to use Ellwood mmf gauge measurements to determine the core reluctance, but these may be omitted and the method of Fig. 9 employed, provided saturation is confined to the core. Subject to this same limitation, the design constants may be evaluated from the pull measurements if these are supplemented with mmf gauge measurements.

For armature saturation, the magnetic design circuit constants can only be evaluated from magnetization measurements, which must include armature search coil as well as coil flux determinations, and be supplemented by mmf gauge measurements.

Limits of Application

The validity and usefulness of the procedures described here rest on the agreement of measured quantities with the relations used to analyse them. As all the relations used lead to linear plots, the extent of agreement in any specific case is apparent from the plot obtained. So far as the presentation of pull results is concerned, this is the only question of validity involved.

The relations found for the magnetization results are used to estimate the field energy and the pull. To the extent the magnetization results relate to the flux linkages of the coil, the conclusions drawn from relations fitting those results are valid. This follows from the energy balance considerations discussed in Section 2. The supplementary measurements, such as those with the mmf gauge or an armature search coil, are in principle only convenient means for determining the relations to which the coil magnetization conforms. The relations found by these means are only valid to the extent of such agreement.

The conformity of magnetization relations to the expressions used here, corresponding to the magnetic circuit schematics, is closest for electromagnets in which the reluctance of the iron parts is small compared with that of the joints and air gaps. This condition is satisfied for most ordinary relays and similar electromagnets in the low density range, where the expressions given here most closely apply. The expressions used for the high density range do not give as close agreement, but provide a satisfactory basis for engineering analysis, particularly when saturation is confined to the core. The treatment is less satisfactory for structures that deviate from these conditions, such as those with armature saturation, or those with long cores of small cross section, where the

iron reluctance is a major component through the full range of operation.

ACKNOWLEDGMENT

The procedures described in this article incorporate contributions proposed or developed by H. N. Wagar, M. A. Logan, Mrs. K. R. Randall, and others.

REFERENCES

1. R. L. Peek, Jr., and H. N. Wagar, Magnetic Design of Relays, page 23 of this issue.
2. A. C. Keller, New General Purpose Relay, B.S.T.J., **31**, p. 1023, Nov., 1952.
3. H. N. Wagar, Relay Measuring Equipment, page 3 of this issue.
4. W. B. Ellwood, A New Magnetomotive Force Gauge, Rev. Sci. Instr., **17**, p. 109, 1946.
5. Estimation and Control of the Operate Time of Relays: Part I — Theory, R. L. Peek, Jr., page 109 of this issue. Part II — Applications, M. A. Logan, page 144 of this issue.
6. Kennelly, Trans. A.I.E.E., **8**, p. 485, 1891.

Estimation and Control of the Operate Time of Relays

Part I—Theory

By R. L. PEEK, Jr.

(Manuscript received August 31, 1953)

The dynamic equations applying to the operation and release of electromagnets are derived in a form in which the armature position and the flux linkages of the coil are the variables to be determined as functions of time. The effective magnetomotive force and pull are taken as given by the static magnetization relations. This formulation is valid if the dynamic field pattern is substantially that of the static field, and it is shown that this condition is satisfied if the effective conductance of the eddy current paths is small compared with that of the coil or other linking circuits. This is usually the case in operation, but is not the case in normal release.

Approximate solutions are obtained for both fast and slow operation, giving the time as related to the mechanical work done, the mass and travel of the moving parts, and the steady state power input to the coil. Expressions are derived for optimum coil design and pole face area, and design requirements for fast operation are discussed.

An approximate expression is derived for the release waiting time, the initial period of field decay prior to armature motion. The effect on this decay of contact protection in the coil circuit is discussed. The slow release case is treated in another article in this issue. An analysis is given of the armature motion in release, and the results of an analog computer solution to this problem reported. These show the effect of design parameters, particularly the armature mass, on the velocity attained in release motion, and hence on the amplitude of armature rebound.

INTRODUCTION

The operate and release times of most telephone relays lie in the range from 1 to 100 millisecs. (0.001 to 0.100 secs.). To use these relays in telephone switching, involving complex patterns of sequential switch

and relay operation, it is necessary: (1) to be able to estimate the operate and release times of any relay for which this information is needed in circuit studies, and (2) to develop relays capable of meeting specific timing requirements, both fast and slow.

The essential basis for such design and estimation is a dynamic theory of the operation of electromagnets, of sufficient accuracy for engineering use. The theory presented here is applicable to electromagnets in general, although the applications discussed are those relating to relays. The theory presented is approximate, partly because of the difficulty of providing a more exact treatment, and partly because simplicity and generality are more important for engineering purposes than accuracy in estimation, which is in any case subject to correction by measured results. The theoretical relations are used alone only in preliminary estimation and in the initial stages of development, as discussed in Part I of this article. In advanced development and in the modification and application of existing structures, as discussed in Part II, the theory is used as a guide in the correlation and extrapolation of observed performance.

The dynamics of electromagnets involve the concurrent and inter-related phenomena of field development and armature motion, and are accordingly governed by the two differential force equations respectively applying, each containing a coupling term expressing their reaction on each other. These basic equations are formally identical with those of electromechanical transducers, such as loud-speakers, but the treatment has little else in common. The operation of an electromagnet is the transient change from one state of equilibrium to another, as distinguished from the sustained low amplitude oscillations of the transducer case.

The coupling terms represent the effect of field energy changes on the coil voltage and on the force causing armature motion respectively. In the steady state, the field energy is a function of magnetomotive force and of armature position alone. In a transient state, eddy currents in the magnetic members affect both field energy and the pattern of the field. In developing an approximate theory, it is assumed that these effects are confined to the total effective mmf, and that the pattern of the field is the same as that of the static field: i.e., that the field energy associated with any portion of the structure, such as an air gap, is fixed by the flux linkages of the coil and by the armature position. The limitations on the analysis imposed by this simplifying assumption are discussed at the end of the next section.

1 THE DYNAMIC EQUATIONS

The notation of this article conforms to the list given on page 257.

THE ELECTRICAL EQUATION

The voltage equation for a coil of N turns linking a field of strength φ may be written in the form:

$$(i - I)R + N \frac{d\varphi}{dt} = 0, \quad (1)$$

where i is the instantaneous current, R is the circuit resistance, and IR is the constant voltage applied. Writing $\mathfrak{F}_i$ for the instantaneous mmf $4\pi Ni$, and $\mathfrak{F}_s$ for the steady state mmf $4\pi NI$, the equation becomes:

$$\mathfrak{F}_i - \mathfrak{F}_s + 4\pi G_i \frac{d\varphi}{dt} = 0,$$

where $G_i = N^2/R$, the coil constant, or equivalent single turn conductance. If the field φ is linked by a number of circuits, a similar voltage equation applies to each such circuit. By addition of these expressions, there is obtained:

$$\mathfrak{F} - \mathfrak{F}_s + 4\pi G \frac{d\varphi}{dt} = 0,$$

where $\mathfrak{F} = \sum \mathfrak{F}_i$, the total effective mmf, $\mathfrak{F}_s = \sum \mathfrak{F}_{si}$, the total applied mmf, and $G = \sum G_i$, the total equivalent single turn conductance. If the dynamic field has the same pattern as the static field, the instantaneous mmf $\mathfrak{F}$ must equal $\mathfrak{R}\varphi$, where $\mathfrak{R}$ is the reluctance $\mathfrak{F}/\varphi$ observed in static magnetization measurements. The preceding equation may therefore be written in the form:

$$\mathfrak{R}\varphi - \mathfrak{F}_s + 4\pi G \frac{d\varphi}{dt} = 0. \quad (2)$$

This relation controls the time rate of development of the flux φ . As the coil is the only circuit linking the field that has an externally applied voltage, $\mathfrak{F}_s$ is simply the steady state applied mmf $4\pi NI$. The conductance G includes not only the coil conductance G_c , or N^2/R , but terms for all conductive paths linking the field, including the eddy current paths. The effective conductance of the eddy current paths is denoted G_E . When a short circuited winding or sleeve is used, as in slow release relays, its conductance is denoted G_s . Thus in most cases of interest $G = G_c + G_E$; when a sleeve is used, $G = G_c + G_E + G_s$.

The static magnetization relations between $\mathfrak{F}$ and φ vary with the position of the armature, which is defined by its displacement x from the operated position. Hence the reluctance $\mathfrak{R}$ is, in general, a function of

both x and φ . Thus equation (2) is a differential equation in which φ , x , and $d\varphi/dt$ are the variables. With the armature at rest, x is constant, and (2) may be solved to give the initial field development in operate, or decay in release. When the armature is moving, the variation of φ and x is governed jointly by (2) and the mechanical force equation.

THE MECHANICAL EQUATION

The forces acting to accelerate the armature and associated moving parts, of effective mass m , are the magnetic pull F and the forces exerted by the contact and other springs. The total spring force may be written as dV/dx , where V is the potential energy stored in the springs, expressed as a function of x . The force equation may be written in the form:

$$F + m \frac{d^2x}{dt^2} + \frac{dV}{dx} = 0. \quad (3)$$

As x measures the gap opening, the velocity dx/dt is positive as the gap opens, and negative as it closes. The pull is directed toward the closed position and therefore tends to algebraically decrease the velocity. The spring force tends to open the gap and to algebraically increase the velocity, as V decreases with x , and dV/dx is negative.

On the assumption that the dynamic and static field patterns are the same, the pull F can be evaluated from the static magnetization relations. F is a function of φ and x , the variables of the flux development equation (2). Thus the dynamic performance is governed jointly by (2) and (3), to which the concurrent variation in φ and x with time must conform. To complete the formulation there are required explicit expressions for $\mathcal{R}$ and F in terms of φ and x .

RELUCTANCE AND PULL

As shown in a companion article,¹ the magnetization relations for most electromagnets conform to the simple magnetic circuit of Fig. 1. Through the region of linear magnetization, in which the flux density is below that producing incipient saturation, the reluctances $\mathcal{R}_0$ and $\mathcal{R}_L$ and the area A are constants. If the magnetization relations conform to this schematic, the total reluctance $\mathcal{R}$ is given by:

$$\mathcal{R} = \frac{\mathcal{R}_L \left(\mathcal{R}_0 + \frac{x}{A} \right)}{\mathcal{R}_L + \mathcal{R}_0 + \frac{x}{A}}. \quad (4)$$

The constants $\mathcal{R}_0$, $\mathcal{R}_L$, and A of Fig. 1 and equation (4) are called

the equivalent values of the closed gap reluctance, the leakage reluctance, and the pole face area respectively. Procedures for estimating these quantities from the magnet's dimensions and material constants are described in the article cited above,¹ while methods for their experimental evaluation are described in another article in this issue.² When equation (4) applies, the pull F is given by:

$$F = \frac{(\mathfrak{R}_L \varphi)^2}{8\pi A \left(\mathfrak{R}_L + \mathfrak{R}_0 + \frac{x}{A} \right)^2} \quad (5)$$

In normal operation, the flux level lies in the region of linear magnetization, and equations (4) and (5) apply through the greater part of the period of flux development determining the operate time. In what fol-

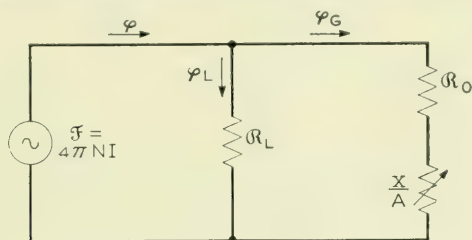


Fig. 1 — Equivalent magnetic circuit of an electromagnet.

lows, some simplification of notation is obtained by writing x_0 for $A\mathfrak{R}_0$, and C_L for $(\mathfrak{R}_0 + \mathfrak{R}_L)/\mathfrak{R}_0$, so that these equations become:

$$\mathfrak{R} = \mathfrak{R}_0 \frac{(C'_L - 1)(x_0 + x)}{C'_L x_0 + x}, \quad (4A)$$

$$F = \left(\frac{(C'_L - 1)x_0}{C'_L x_0 + x} \right)^2 \frac{\varphi^2}{8\pi A}. \quad (5A)$$

Thus the dynamic relations are given by (2) and (3), in which $\mathfrak{R}$ and F are given by (4) and (5) respectively. The magnetic circuit constants can be evaluated by the methods cited above, and the other constant terms are known or given quantities, with the exception of the eddy current conductance G_E . Evaluation of this term requires determination of the conditions under which such a constant term can adequately represent the effects of eddy currents.

EDDY CURRENT CONDUCTANCE

Fig. 2 shows a simple electromagnet with a cylindrical core of length ℓ and diameter D . If $\mathfrak{R}$ is the reluctance, the magnetomotive force $\mathcal{F}_c$ of

the winding current results in a flux $\varphi_c = \mathfrak{F}_c/\mathcal{R}$, uniformly distributed across the core cross section. The leakage paths cause some longitudinal variation in φ , which can be neglected for the present purpose. If the flux is varying, eddy currents flow in circular paths with a current density j , which varies with the radius r . These give an increment to the total mmf varying from zero at the surface to a maximum at the center, producing a corresponding variation in the density of the total flux. If $\mathfrak{F}_c$ is large compared with the maximum magnetomotive force produced by the eddy currents, however, the density is nearly constant, and its rate of change is nearly constant throughout the core cross section. To the extent that this condition is satisfied, the effect of the eddy currents can be determined as follows.

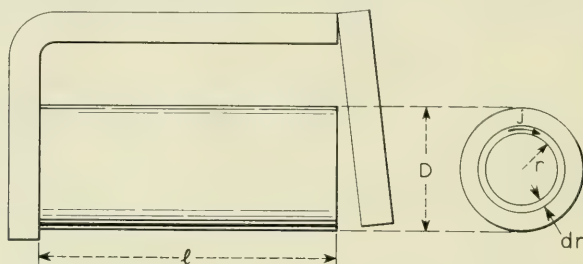


Fig. 2 — Eddy current paths in the core of an electromagnet.

The current in a cylindrical shell of radius r and differential thickness dr is $j\ell dr$. This links a part of the field proportional to the area enclosed, or $4r^2\varphi/D^2$, which varies, by hypothesis, at the same proportional rate as the total field. The resistance of the shell is $2\pi r\rho/(\ell dr)$, where ρ is the resistivity of the material. The voltage equation for the shell circuit is therefore:

$$j\ell dr \frac{2\pi r\rho}{\ell dr} + \frac{4r^2}{D^2} \frac{d\varphi}{dt} = 0,$$

and hence:

$$j = -\frac{2r}{\rho\pi D^2} \frac{d\varphi}{dt}.$$

The magnetomotive force of the shell is its current multiplied by 4π . This produces a flux increment $d\varphi$ in the area enclosed inversely proportional to the reluctance of the tubes of induction within this area. This reluctance may be taken as inversely proportional to the area, as would be the case in a closed magnetic circuit of uniform cross section. Then

the flux increment produced by each shell is given by:

$$d\varphi = 4\pi j \ell dr \frac{4r^2}{D^2(2\ell)}.$$

On substituting the preceding expression for the current density j , and evaluating the integral $\int_0^{D/2} d\varphi$, the total flux φ_E produced by the eddy currents is found to be:

$$\varphi_E = -\frac{\ell}{2\rho 2\ell} \frac{d\varphi}{dt}. \quad (6)$$

This is identical in form with the expression for one of the linking circuits assumed in deriving equation (2), with $2\ell\varphi_E$ equal to $\mathfrak{F}_E$, and $4\pi G_E$ given by $\ell/(2\rho)$. Thus to the extent that the assumption of a uniform flux distribution is valid, G_E is given by:

$$G_E = \frac{\ell}{8\pi\rho}. \quad (7)$$

This expression applies to a round core. A parallel approximation may be obtained for a rectangular rod by considering it as made up of rectangular shells of differential thickness, having their sides in the same ratio as the rod. By treating the perimeter of each shell as corresponding to the circumference of the circular shell in the cylinder case, with the area enclosed corresponding to that enclosed by the circular shell, the following expression is obtained for the conductance of a rectangular rod whose sides are in the ratio k :

$$G_E = \frac{\ell}{8\pi\rho} \frac{\sqrt{\pi k}}{1+k}. \quad (8)$$

This shows that the effect of using a rectangular section is equivalent to increasing the resistivity of the material. A similar effect is obtained by subdividing the core section, as in laminated construction. If the core, for example, were made of two similar round rods, equation (6) would apply to each, and the effective mmf acting on both would be the value of $2\ell\varphi_E$ given by this expression, with φ in $d\varphi/dt$ equal to half the total flux, so that the resulting expression for G_E would be half that given by (6). This argument can be generalized to show that, aside from the effect of changes in section shape, the effect of subdivision is to reduce G_E in proportion to the number of subdivisions.

The approximate validity of these expressions for G_E rests on the assumption that the eddy current magnetomotive force is minor com-

pared with that of the coil or other external circuit. From the derivation of (2), the component mmf's are proportional to the corresponding terms in G , so the preceding expressions for G_E are valid only if G_E is small compared with $G_c(N^2/R)$. In all but exceptional cases, this condition is satisfied in operation, and in release with a short circuited winding or a slow release sleeve. In these cases, the effect of eddy currents is adequately represented by a constant G_E term in (2). The exact values of G_E applying are in only approximate agreement with those given by (7) and (8), as the derivation of these expressions ignores the variation in flux density along the length of the core, and the eddy current effects in the armature and return path.

EDDY CURRENTS IN RELEASE

In normal release, the winding circuit is open, and the only effective magnetomotive force is that of the eddy currents. Evidently, the field in the outer layers must collapse almost instantly, while that in the center of the core is sustained by eddy currents in paths whose mean conductance, per line linked, is higher than that applying to a uniform field.

The relations applying to a closed path of uniform section can be formulated in differential form, and the solution for a cylindrical section has been given by Wwedensky.³ For decay from an initially uniform field, the expression for the flux as a function of time is a series of exponential terms with progressively smaller time constants. The first term represents the most persistent part of the flux, and represents a field varying from zero at the surface to a maximum at the center, comprising 69 per cent of the initial field. Its time constant corresponds to an effective value of G_E 35 per cent larger than that given by (7). Somewhat similar relations must apply to an electromagnet, causing a time variation in the pattern of the field, not only radially, as in a closed uniform path, but in the longitudinal variation and in the division of the field between the leakage and armature paths.

An experimental study of flux development and decay in relays has been reported by M. A. Logan.⁴ His results agree with this discussion in showing G_E in (2) to be effectively a constant, provided G_E/G_c is less than 0.2, as in the operation of most relays. An empirical expression is given for an effective value of $G_c + G_E$ which provides a correction factor applicable for small values of G_E/G_c . The results for normal release show the field decay to have the general character of Wwedensky's solution, and an empirical expression is given which agrees with observed results. This is primarily of interest in connection with the voltage

induced in the coil and imposed on the contact opening the coil circuit. Because of the changing field pattern, the gap field which determines the pull has a different, though similar, decay rate.

These considerations, as supported by Wwedensky's relations and Logan's results, indicate that the rate of pull decay in normal release is faster, and only roughly of the same order of magnitude, as that estimated on the assumption that (2) applies, with values of G_E similar to those applying in operation.

LIMITATIONS OF THE ANALYSIS

The validity of equations (2) through (5) rests on the assumption that the pattern of the dynamic field is essentially that of the static field to which the magnetization relations apply. The above discussion of the eddy currents indicates that this assumption is valid when G_E is a minor term in G . This condition is approximately satisfied in normal operation. In open circuit release, however, the condition is not satisfied, and equation (2) is only a crude first approximation to the controlling relation.

The use of (4) as an expression for the reluctance $\mathfrak{R}$ rests on the further assumption that the magnetization is linear, a condition only satisfied in the low density region. The initial and controlling stages of operation are usually complete before the field passes out of the low density region, and (4) is therefore applicable to the operate case. A different expression for $\mathfrak{R}$ is required in the release case, as discussed in Section 7.

2 CHARACTER OF THE OPERATE SOLUTION

GRAPHICAL REPRESENTATION

Some understanding of the relations applying to operation may be obtained from their graphical representation in Figs. 3 and 4. In these two figures the path followed by the variables in dynamic operation is indicated by the dotted lines, with the dots spaced to indicate equal time intervals between them. Fig. 3 shows the φ versus $\bar{x}$ relation, referred to the steady state magnetization curves for various values of x , with $x = 0$ corresponding to the operated position, and $x = x_1$ to the initial unoperated position. Fig. 4 shows the dynamic F versus x relation, together with the load curve (bounding the cross hatched area V) and the steady state pull curve for the applied mmf $\mathfrak{F}_s$.

The flux and pull increase together with the armature at rest at x_1 until the pull equals the back tension at the point 1. In the earlier motion, 1-2, the velocity is small, and the reluctance (from (4)) changes slowly with x so the motion has little effect on the rate of flux develop-

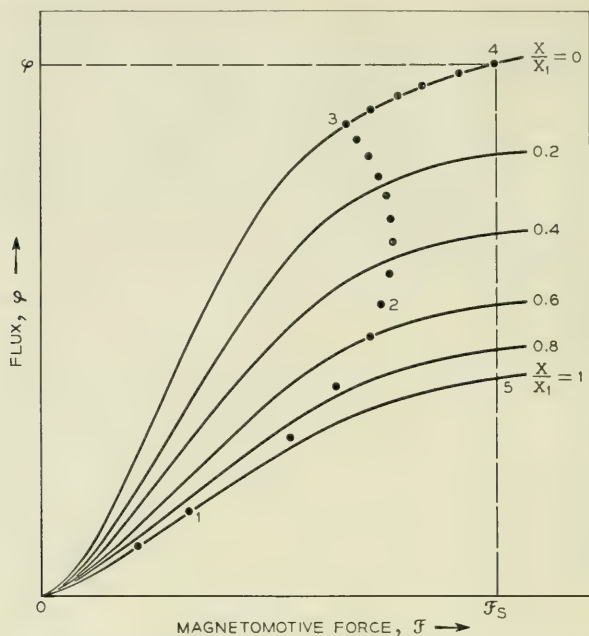


Fig. 3 — Field energy relations in the operation of an electromagnet.

ment. In the later motion, 2-3, the reluctance changes more rapidly with x , and the velocity is high, increasing $d\phi/dt$ so as to result in a temporary decrease in $\mathcal{F}$. Operation is complete at 3, with ϕ and F still below their steady state values at 4, which they then approach exponentially with $x = 0$.

The mechanical work done in operation is represented in Fig. 4 by the area under the dynamic pull curve (dotted line), and in Fig. 3 by the area bounded by 0-1-2-3-0. This, of course, is less than the work that would be done if $\mathcal{F}$ were equal to its steady state value $\mathcal{F}_s$ throughout the motion, represented by the area under the $\mathcal{F}_s$ curve in Fig. 4, and by the loop 0-3-4-5-0 in Fig. 3. In Fig. 4 it can be seen that the work done exceeds the static load V by an amount represented by the shaded area T , corresponding to the kinetic energy of the armature and the parts that move with it. At the end of the stroke this energy is partly dissipated in impact, and partly transferred to vibratory motion of the relay and its parts.

The initial flux development, 0-1 is governed wholly by equation (2). This same relation dominates in controlling the performance in the early travel 1-2, in which the velocity is small, and the reluctance changes slowly. In the later travel, where the velocity is high and the reluctance

changes rapidly, equation (3) is controlling, with F corresponding to a more nearly constant value of φ .

METHODS OF SOLUTION

Formally, the expression for φ given by combining (3) and (5) may be substituted in (2), eliminating φ and giving a third order non-linear differential equation in x and t . No analytical solution to this equation is known, even for the idealized case of a linear load curve. Specific solutions may be obtained by means of computing devices, but there are too many parameters involved to prepare a useful library of specific solutions that would serve as a general treatment.

Approximate analytical solutions can be obtained by a step by step process, starting from the initial stage of flux development with the armature at rest. These solutions are relatively simple in form and application, and provide an analytical statement of the relations between performance and design parameters, from which conclusions can be drawn as to design features and operating conditions favoring a desired



Fig. 4 Load and pull relations in the operation of an electromagnet.

type of performance. For development purposes, an approximate solution indicating such optimum conditions is preferred to a rigorous solution in a form too complex for such use.

Three such approximate solutions for the operate case are described below. The first of these, the single stage approximation, is an application of the solution to the flux rise equation (2), neglecting the armature motion. It is applicable in cases of slow operation controlled by the rate at which the flux development provides sufficient pull to operate the load. The other two solutions are more generally applicable, particularly for relatively fast operation. One of these, the two-stage approximation, is the simpler in form, while the other, the three stage approximation, is the more accurate.

3 SINGLE STAGE APPROXIMATION

The initial stage of flux development with the armature at rest ends when the pull equals the back tension. During this stage the flux development is governed wholly by (2), with $x = x_1$, the initial gap, and $\mathcal{R} = \mathcal{R}_1$, as given by (4) for $x = x_1$. Writing $\varphi_1\mathcal{R}_1$ for $\mathfrak{F}_s$, equation (2) becomes:

$$t = \frac{4\pi G}{\mathcal{R}_1} \int_0^\varphi \frac{d\varphi}{\varphi_1 - \varphi},$$

which on integration gives:

$$t = \frac{4\pi G}{\mathcal{R}_1} \ln \frac{1}{1 - v}, \quad (9)$$

where $v = \varphi/\varphi_1$, the ratio of the flux attained at time t to the steady state flux φ_1 or $\mathfrak{F}_s/\mathcal{R}_1$. The mmf ratio $\mathfrak{F}/\mathfrak{F}_s$ is also equal to v .

Equation (9) applies rigorously only while the armature is at rest. It is, however, also a close approximation for the initial motion, in which the velocity is small and x differs little from x_1 . If the rate of flux development is slow ($G/\mathcal{R}_1$ large) the armature moves slowly with the pull only slightly in excess of the load curve. In this case the inertia term $m \cdot d^2x/dt^2$ in (3) is minor, and the pull F nearly equals the static load dV/dx . Thus the operate time is approximately equal to the time required to develop a pull which exceeds the static load at all points in the travel. This pull is attained at the just operate current or ampere turn value, corresponding to the minimum mmf for static operation, $\mathfrak{F}_0$. Assuming that the armature moves as the flux and pull develop, the operate time is that required for $\mathfrak{F}$ to equal $\mathfrak{F}_0$, and is given by equation

(9) for $v = \mathfrak{F}_0/\mathfrak{F}_s$. While this is a crude approximation, it provides satisfactory estimates of the time for slow operation.

EDDY CURRENT CONDUCTANCE DETERMINATION

Equation (9) is used in one experimental method for the evaluation of the eddy current conductance G_E . In this, measurements are made of the time at which motion starts for various values of the coil constant G_C . The latter may be varied by adding series resistance, and the applied voltage adjusted to maintain the current and hence $\mathfrak{F}_s$ constant in successive measurements. The time t for motion to start may be determined from shadowgraph measurements. For a constant back tension, the pull and hence the value of v or $\mathfrak{F}/\mathfrak{F}_s$ when motion starts is constant in these measurements. From (9), t is then directly proportional to $G_C + G_E$, so that a plot of t vs. G_C should be linear, with a negative intercept on the G_C axis numerically equal to G_E . Experimentally, an approximately linear relation is obtained in this way, provided the values of G_C covered are in excess of $5G_E$, in agreement with the discussion of Section 1. Values of G_E thus determined are consistent with those obtained from other measurements, and in approximate agreement with estimates obtained from (7) or (8).

COIL CONSTANT

In equation (9) G_E and $\mathfrak{R}_1$ are determined by the magnetic design. The latter, together with the load, determines $\mathfrak{F}_0$, the just operate mmf, or $4\pi(NI)_0$. As v equals $\mathfrak{F}_0/\mathfrak{F}_s$, or $(NI)_0/(NI)$, the only quantities not fixed by the design and the load requirement are G_C and the steady state ampere turn value NI . As the square of this latter quantity is equal to $G_C \cdot I^2 R$, equation (9) may be written in the form:

$$t_0 = \frac{4\pi}{\mathfrak{R}_1} \left(G_E + \frac{(NI)_0^2}{v^2 I^2 R} \right) \ln \frac{1}{1-v}. \quad (10)$$

The steady state power $I^2 R$ is either determined by circuit requirements, or chosen with reference to economy of power consumption. With $I^2 R$ fixed, the only independent quantity in (10) is v , which is determined by the choice of the coil constant G_C . Neglecting G_E , the time is proportional to

$$\frac{1}{v^2} \ln \frac{1}{1-v},$$

which is shown plotted against v in Fig. 5. This figure includes a curve

giving corresponding values of $\ln 1/(1 - v)$, which may be used in specific cases to determine the correction corresponding to the term in G_E .

In most cases of slow operation, the operate time is of little importance, and G_c is chosen to make NI greater than $(NI)_0$ in all cases: i.e., after allowing for possible variations in $(NI)_0$ resulting from variations in the load and in the magnetic characteristics. These variations result in a variation in v , and the corresponding variation in time is shown by the curve of Fig. 5. If, however, it is desired to minimize the operate time for a given power input, G_c should be chosen to give the value of v (0.715) corresponding to the minimum of the curve. As this minimum is broad, variations in v , corresponding to those in $(NI)_0$, produce little change in the operate time.

4 TWO STAGE APPROXIMATION

Unless the rate of flux development is very slow, as assumed in the single stage approximation, the pull attained in the early travel is in excess of the load, and the kinetic energy T is a considerable part of the total work output $V + T$. As a first approximation to this case, the operate time can be computed as though operation occurred in two successive stages: (1) a stage of flux development with the armature at rest in the unoperated position ($x = x_1$), and (2) a stage of motion, in which

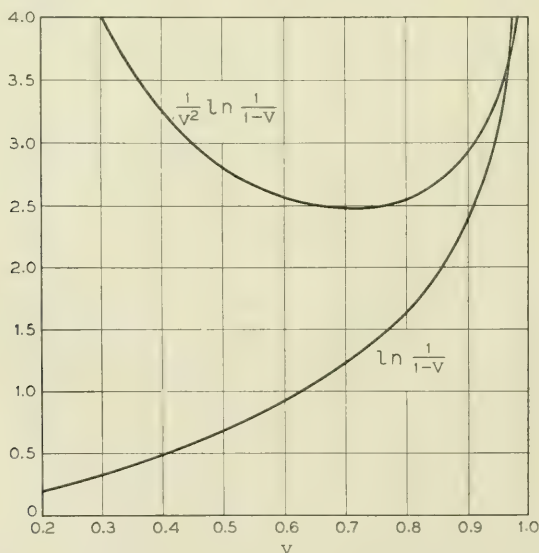


Fig. 5 — Relations for estimating time of flux development.

the flux remains constant at the value attained at the end of the first stage. This approximation necessarily over-estimates the operate time, as it takes the two stages as occurring in sequence, whereas they actually overlap.

Let t_1 be the time of the first stage, and t_2 the time of the second stage, where t_2 is the time for the motion from x_1 to some other armature position x_2 , measuring the completion of operation. The first stage ends when the flux has attained some value $v\varphi_1$, where $\varphi_1 = \mathfrak{F}_s/\mathfrak{R}_1$, the steady state flux for the open gap, and v remains to be determined. Then t_1 is given by (9) for this value of v .

By hypothesis, the flux remains constant at $v\varphi_1$ throughout the second stage, and the corresponding pull is then given by (5A) for $\varphi = v\varphi_1$. The work $V + T$ is the integral of $F \cdot dx$ from x_1 to x_2 , given by the integral of (5A). Hence:

$$V + T = \frac{(C_L - 1)v^2\varphi_1^2}{8\pi A} \frac{x_0^2(x_1 - x_2)}{(C_L x_0 + x_1)(C_L x_0 + x_2)}. \quad (11)$$

In this equation x_0/A may be replaced by the expression for $\mathfrak{R}_0$ given by (4A) for $x = x_1$, and $G_c \cdot I^2 R$ substituted for $(NI)^2$ in the resulting equation, which then becomes:

$$v^2 = \frac{\mathfrak{R}_1 C_w}{2\pi G_c I^2 R} (V + T), \quad (12)$$

where C_w is given by:

$$C_w = \frac{(x_0 + x_1)(C_L x_0 + x_2)}{(C_L - 1)x_0(x_1 - x_2)}. \quad (13)$$

The time t_2 of the second stage depends upon the difference between the pull and load curves. As a representative condition, this may be taken as approximately constant. For such uniform acceleration of the effective mass m through the distance $x_1 - x_2$, the kinetic energy T equals $2m(x_1 - x_2)^2/t_2^2$, giving an expression for t_2 in terms of T . As the time t_1 of the first stage is given by (9), the total operate time t_0 , or $t_1 + t_2$ is given by:

$$t_0 = \frac{4\pi(G_E + G_c)}{\mathfrak{R}_1} \ln \frac{1}{1 - v} + \left(\frac{2m(x_1 - x_2)^2}{T} \right)^{\frac{1}{2}}. \quad (14)$$

In this approximate expression for the operate time, v and T are as yet undetermined quantities, related by equation (12), in which G_c is the value of N^2/R for the coil used. The two stage operation assumed in deriving this expression corresponds to the existence of a restraint

which holds the armature at rest until it attains the flux level $v\phi_1$. The approximation will most nearly approach the true relation if the effect of this restraint is minimized, and v taken at the value which minimizes t_0 . This minimum value can be determined by trial, using the curves of Fig. 5 to compute t_0 for several values of v , and taking the minimum from the resulting plot of t_0 versus v .

In development studies, the operate time of interest is usually not that for an assigned winding, but that for the winding giving minimum time for a given power input. Substituting in (14) the expression for G_c given by (12), there is obtained the equation:

$$t_0 = \left(t_E + \frac{2C_w}{v^2} \frac{V}{I^2 R} \right) \ln \frac{1}{1-v} + \left(\frac{2m(x_1 - x_2)^2}{T} \right)^{\frac{1}{2}},$$

where t_E is written for $4\pi G_E/\mathcal{R}_1$, the eddy current time constant. As G_c remains to be determined, this expression contains two independent variables, v and T , which are to be so chosen as to make t_0 a minimum. Equating the partial derivative with respect to T to zero, there is obtained the following expression for the value of T for which the time is a minimum:

$$T = 2m(x_1 - x_2)^2 \left(\frac{v^2 I^2 R}{4C_w \ln \frac{1}{1-v}} \right)^2.$$

On substituting this expression for T , the preceding equation becomes:

$$t_0 = \left(t_E + \frac{2C_w}{v^2} \frac{V}{I^2 R} \right) \ln \frac{1}{1-v} + \left(\frac{27C_w}{v^2} \ln \frac{1}{1-v} \frac{m(x_1 - x_2)^2}{I^2 R} \right)^{\frac{1}{2}}. \quad (15)$$

If the value of v which minimizes t_0 is determined, the resulting value of t_0 is the minimum operate time attainable by optimum coil design. In particular, if t_E is negligible, this minimum corresponds, from Fig. 5, to $v = 0.715$, for which

$$\frac{1}{v^2} \ln \frac{1}{1-v} = 2.5.$$

The corresponding value of G_c , the coil constant value making t_0 a minimum, is given by (12). In this use of equation (12), v and T are taken at the optimum values obtained as indicated above.

OPTIMUM COIL CONSTANT

As in the case of the single stage approximation, the optimum coil constant corresponds to a value for v of 0.715 when t_E is negligible. From

the curves of Fig. 5, it can be seen that if t_E is not negligible, the optimum value of r is less than 0.715, and will be smaller the larger the ratio of the first term in (15), that in t_E , to the other two terms. From (12), a reduction in r corresponds to an increase in NI , and hence in the coil constant G_C (I^2R being given). Thus the optimum coil constant for fast operation is larger for electromagnets with long cores of low resistivity material, for which G_E and t_E are large, than for those with short cores of high resistivity material, for which these quantities are small.

The physical significance of the optimum coil constant can be deduced from the relations shown graphically in Fig. 4. For a given value of the power input I^2R , an increase in G_C increases both the steady state mmf $\mathfrak{F}_s$ and the time constant of flux development $4\pi G_C/\mathfrak{R}_1$, increasing the upper limit to the attainable pull while reducing the rate at which this limit is approached. If the coil constant were so small that the area under the dynamic pull curve equalled the static work load V , the operate time would be infinite. On the other hand, an infinitely large coil constant would correspond to infinite inductance and an infinite operate time. Between these extremes lies the optimum value of the coil constant, giving minimum operate time. This is a broad optimum, near which a change in $\mathfrak{F}_s$ is compensated by a corresponding change in the rate of flux development, so that the realized dynamic pull curve is not affected by a small change in G_C .

In addition to the operate time, the value of G_C affects the final velocity of the armature, and thus its kinetic energy in impact with the core at the end of the stroke. This energy is dissipated in the relay and spring vibration associated with contact chatter. For values of G_C above the optimum, the higher inductance reduces the pull in the early travel, while the higher value of $\mathfrak{F}_s$ increases it in the later travel. The net effect is to increase both the operate time and the final velocity. Values of G_C above the optimum are therefore disadvantageous not only in slower operation, but in increased impact energy tending to cause contact chatter.

FACTORS CONTROLLING SPEED

Aside from the term in t_E , equation (15) shows that the minimum operate time, corresponding to the optimum value of r , is determined by the static load V , the inertia load measured by $m(x_1 - x_2)^2$, the steady state power I^2R , and the constant C_W . The latter is given by (13), and is the only quantity in (15), aside from t_E , which depends upon the magnetic design. In the range of values applying in practice, C_W is determined primarily by the leakage factor C_L , measuring the ratio of leakage to useful flux. In most practical cases, x_2 is small (zero for complete

operation), C_L equals or exceeds 4 ($\mathfrak{G}_L$ equals or exceeds 3 $\mathfrak{G}_0$), and x_1/x_0 lies in the range from 1 to 3. For $4 < C_L < 10$ and $1 < x_1/x_0 < 3$, $1.5 < C_w < 2.6$. In practice therefore C_w has a value close to 2.0 for most electromagnets.

The term in V , the second term in (15), varies inversely as the power I^2R , while the third, or inertia term, varies as the cube root of the power. Hence the second term tends to dominate at low values of I^2R , and the third term tends to dominate at high values. For a low power input therefore the operate time is controlled by the load V , and varies inversely as the power input, while for a high power input, the operate time is controlled by the inertia term, $m(x_1 - x_2)^2$, and varies inversely as the cube root of the power. In the load controlled case, the time varies directly as the load; in the mass controlled case it varies as the cube root of the mass and as the two-thirds power of the travel $x_1 - x_2$. The eddy current term is very nearly a constant increment to the other two terms, and is therefore relatively more important, the faster the operation.

PRELIMINARY TIMING ESTIMATES

Within the limits of accuracy to which this two stage approximation applies, the effect of the magnet design upon the operate time appears only in t_E and C_w . If the former is neglected, the optimum value of v is 0.715, and

$$\frac{1}{v^2} \ln \frac{1}{1-v} = 2.5.$$

Taking C_w as having the representative value of 2.0, (15) reduces to:

$$t_0 = \frac{10V}{I^2R} + \left(\frac{135m(x_1 - x_2)^2}{I^2R} \right)^{\frac{1}{3}}. \quad (16)$$

This simple approximation provides rough estimates of the operate times attainable with any electromagnet for a given load and power input. As an illustration, consider a typical relay spring load involving a travel of 40 mil-in (0.1 cm) with a level of spring force of 100 gm (10^5 dynes), so that $V = 10^4$ ergs. Let the effective mass m of the moving parts be 10 gm. For a steady state power input of 0.5 watts (5×10^6 ergs sec.), the two terms of (16) have values of 0.020 sec and 0.014 sec, respectively, so that t_0 equals 0.034 sec. For an input of 5 watts, the two terms become 0.002 sec and 0.007 sec, respectively, and t_0 equals 0.009 sec. The neglected t_E term of (15) might amount to an increment of 0.005 sec for a solid core of magnetic iron, but would be proportionally smaller for higher

resistivity material. In either case, this increment would be of little consequence in the low power case, but would materially affect the operate time in the high power case.

5 THREE STAGE APPROXIMATION

The following analysis provides greater accuracy in estimating operate time than the two stage approximation, together with a more accurate representation of the relations between performance and the controlling variables. It differs from the two stage approximation in treating the initial motion as a separate stage of operation. The three stages of operation thus become: (1) Increase of the flux to the value at which the pull equals the back tension, with the armature at rest, (2) flux development and concurrent armature acceleration, assuming equations (2) and (3) to apply, with the approximation that the variations of $\mathfrak{R}$ and $\mathfrak{F}$ with x are ignored, and (3) the later motion, which is treated in the same manner as the second stage of the two stage approximation.

Formally, the second stage should be restricted to a very small part of the total armature motion, in order to minimize the change in $\mathfrak{R}$ which is ignored. Relatively minor error, however, is introduced in ignoring the variation in $\mathfrak{R}$ through as much as half the total travel. The spring force is taken as constant through the second stage. In the relay case, this is approximately true for the initial travel prior to the actuation of the contacts, which results in an abrupt increase in the spring load. Thus the second stage can be taken to extend to the travel at which contact actuation first occurs. If all contacts are actuated near the same point in the travel, the end of the second stage coincides with complete operation, and the third stage need not be considered. If the contacts are spread out, the third stage coincides with the stagger time, or time between operation of the first and last contacts. The operate time therefore consists either entirely or principally of the time for the first two stages.

As before let:

x_1 be the initial (open) gap, for which $\mathfrak{R} = \mathfrak{R}_1$,

F_1 be the back tension (constant load in the first two stages),

t_c be the initial coil time constant, $\frac{4\pi}{\mathfrak{R}_1} \cdot \frac{N^2}{R}$,

t_E be the initial eddy current time constant, $\frac{4\pi}{\mathfrak{R}_1} G_E$,

$\mathfrak{F}_s$ be the steady state mmf, $4\pi NI$, and φ_1 the corresponding flux for $x = x_1$, or $\mathfrak{F}_s/\mathfrak{R}_1$.

FIRST STAGE

The pull for $x = x_1$ is given by (5A), which may be written:

$$F = \frac{(v\varphi_1)^2}{8\pi A_1}, \quad (17)$$

where A_1 is given by:

$$A_1 = \left(\frac{C_L x_0 + x_1}{x_0(C_L - 1)} \right)^2 A. \quad (18)$$

In the first stage the flux increases to $v_1\varphi_1$, where v_1 is given by (17) for $F = F_1$. Then the time t_1 of the first stage is given by (9) for $v = v_1$, or by:

$$t_1 = (t_c + t_E) \ln \frac{1}{1 - v_1}. \quad (19)$$

The solution for the first stage therefore determines v_1 and t_1 .

SECOND STAGE

It is convenient to write the equations for the second stage in terms of the ratio z defined by:

$$z = \frac{t}{t_c + t_E},$$

where t is now measured from the start of the second stage. Neglecting the change in reluctance with armature motion, the ratio φ/φ_1 or v is given by (9), except that for $v = v_1$ at $t = 0$ the relation must be written:

$$v = 1 - (1 - v_1)e^{-z}. \quad (20)$$

$$\text{As } \varphi_1 = \mathfrak{F}_s/\mathfrak{R}_1, \quad \text{or} \quad 4\pi NI/\mathfrak{R}_1, \quad \text{and} \quad t_c = \frac{4\pi}{\mathfrak{R}_1} \cdot \frac{N^2}{R},$$

$$\varphi_1^2 = \frac{4\pi}{\mathfrak{R}_1} I^2 R \cdot t_c.$$

Taking the pull in the second stage as given by the expression applying for $x = x_1$, and substituting the preceding expressions for v and φ_1 , there is obtained:

$$F = \frac{I^2 R t_c}{2\mathfrak{R}_1 A_1} (1 - (1 - v_1)e^{-z})^2. \quad (21)$$

For the second stage dV/dx is constant at $-F_1$ throughout, and (3) reduces to $m \, d^2x/dt^2 = F - F_1$, or:

$$\frac{d^2x}{dt^2} = \frac{I^2 R t_c}{2m\phi_1 A_1} ((1 - (1 - v_1)e^{-z})^2 - v_1^2).$$

Substituting $(t_c + t_E)z$ for t , this equation may be integrated with respect to t to obtain the following expressions for the velocity $\dot{x}$ and the travel x in the second stage:

$$\dot{x} = \frac{I^2 R t_c (t_c + t_E)}{2m\phi_1 A_1} f_1(v_1, z), \quad (22)$$

$$x_1 - x = \frac{I^2 R t_c (t_c + t_E)^2}{2m\phi_1 A_1} f_2(v_1, z). \quad (23)$$

The functions $f_1(v_1, z)$ and $f_2(v_1, z)$ are the integrals with respect to z of the bracketed term in the preceding equation. These functions can be evaluated from the curves of Fig. 6, which shows them plotted against z for several values of v_1 .

For a specific case, the values of v_1 and t_1 will have been determined from the relations for the first stage. Let x_2 be the travel at the end of the second stage. Then the value of $f_2(v_1, z)$ for $x = x_2$ is given by (23), and z_2 , the corresponding value of z , can be read from the curve of Fig. 6 for the value of v_1 applying. Then the time for the second stage, t_2 , is given by $z_2(t_c + t_E)$. For $z = z_2$, (20) gives v_2 , the value of v at the end of the second stage. This in turn gives the corresponding value of the flux $v_2\phi_1$, and, from (21), the pull F at the end of this stage. The velocity at the end of this stage is given by (22), with $f_1(v_1, z)$ read from the curves of Fig. 6 for $z = z_2$.

THIRD STAGE

The total operate time is $t_1 + t_2 + t_3$, where t_3 is the time for the third stage. In the relay case, this is the stagger time between the operation of the first and last contacts. A first, and frequently adequate approximation to t_3 , is given by assuming the velocity in the third stage to be constant at the value attained at the end of the second stage. For a more exact determination, particularly in determining the final velocity for complete operation, it may be assumed that the flux in the third stage is constant at the value $v_2\phi_1$ attained at the end of the second stage. The mechanical output $V + T$, in the third stage is then given by equation (11), with v taken as v_2 , and x_1 and x_2 replaced, respectively, by x_2 and x_3 , the latter denoting the travel at the end of the third stage (zero for complete operation). Knowing the spring load V between x_2 and x_3 , t_3 and the final velocity can be computed from the increase in kinetic energy T on the assumption of uniform acceleration.

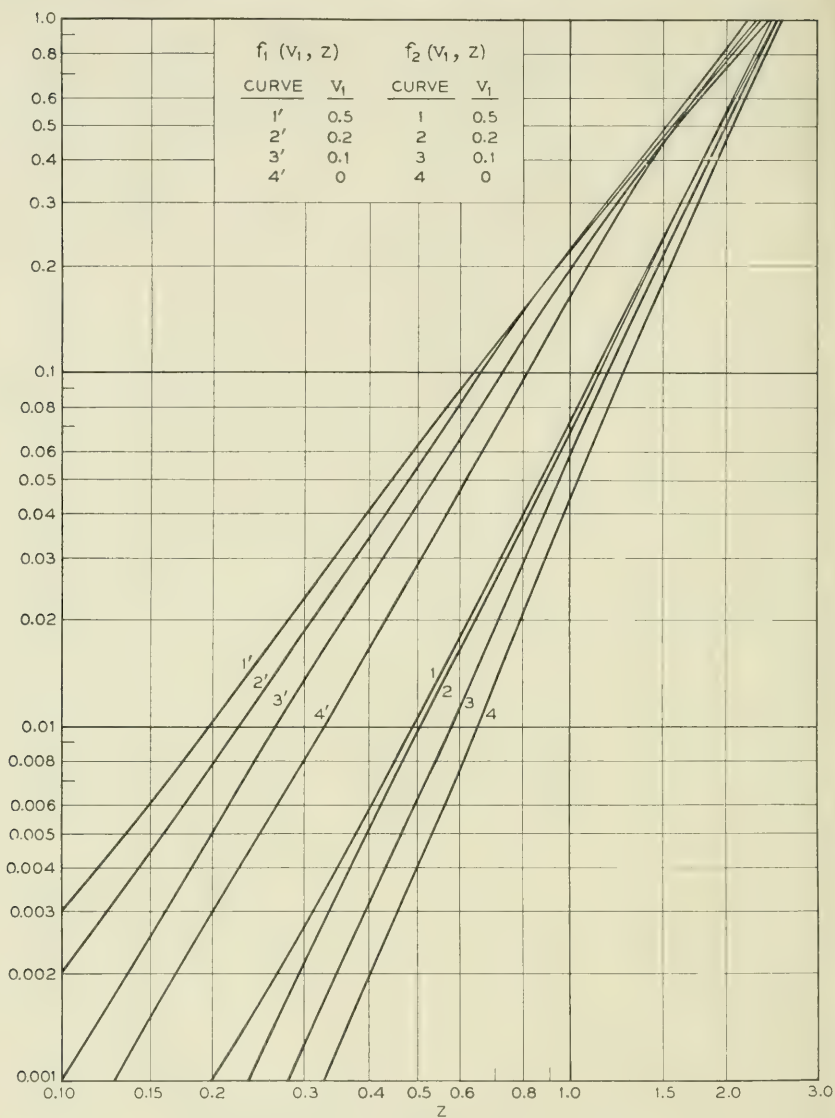


Fig. 6 — Relations for initial motion in operation.

FACTORS AFFECTING OPERATE PERFORMANCE

The dominant term in the operate time is the time t_2 for the second stage, which also determines the velocity and pull level in the third stage, and hence the final velocity. As $t_2 = z_2(t_C + t_E)$, equation (23) gives the following expression for t_2 :

$$t_2^3 = \frac{2mA_1\mathfrak{R}_1(x_1 - x_2)}{I^2R} \frac{t_c + t_k}{t_c} \frac{z_2^3}{f_2(v_1, z_2)}. \quad (24)$$

As can be seen in Fig. 6, $f_2(v_1, z)$ is, to a rough first approximation, proportional to z^3 . To the extent that this approximation holds, t_2 is proportional to the cube root of $m A_1 \mathfrak{R}_1 (x_1 - x_2)$, $(I^2 R)$, as for the last term in (15) and (16), except that $(x_1 - x_2)^2$ is replaced by $A_1 \mathfrak{R}_1 (x_1 - x_2)$. From (4A) and (18), $A_1 \mathfrak{R}_1$ is given by:

$$A_1 \mathfrak{R}_1 = \frac{(x_0 + x_1)(C_L x_0 + x_1)}{(C' - 1)x_0}.$$

Here $x_0 = A \mathfrak{R}_0$, and t_2 is a minimum for that pole face area A for which $A_1 \mathfrak{R}_1$ is a minimum. This condition, as determined by equating the derivative of the preceding expression with respect to x_0 to zero, is given by:

$$x_1^2 = C_L x_0^2,$$

or $x_1 A = \sqrt{C_L} \cdot \mathfrak{R}_0$, where A is the optimum pole face area for fast operation. This is a smaller pole face area than that for maximum sensitivity, for which $x_1 A$ should equal $\mathfrak{R}_0$, as shown in the companion article¹ cited above. If the pole face area satisfies the preceding condition, the expression for A_1 reduces to:

$$A_1 = \frac{\sqrt{C_L} + 1}{\sqrt{C_L} - 1} x_1,$$

and the term $A_1 \mathfrak{R}_1$ in (24) reduces to $x_1(x_1 - x_2)$ multiplied by the factor $(\sqrt{C_L} + 1)/(\sqrt{C_L} - 1)$. This leakage factor, which increases with the ratio of leakage to useful flux, is then the only term in (24) which varies with the design of the electromagnet.

If $f_2(v_1, z)$ were strictly proportional to z^3 , equation (24) would be nearly independent of the coil constant N^2/R , to which t_c is proportional. As $f_2(v_1, z)$ departs from this cube law relation, and also varies with v_1 , t_2 is not independent of the coil constant. The optimum coil constant is that which minimizes $t_1 + t_2$. In any specific case, a succession of values may be assumed for G_C , and t_1 and t_2 evaluated by the relations given above. The resulting relation between $t_1 + t_2$ and G_C is similar in character to that described above in connection with the two stage approximation.

6 DESIGN FOR FAST OPERATION

The approximations discussed above provide a means for estimating the operate time attainable in specific cases, and for selecting certain

design variables to obtain maximum speed. In particular, it has been shown that for a given relay, with a specified load and power input, a winding can be selected to minimize the operate time. Aside from this, the relations show that for a given relay, the operate time depends only upon the load and travel and the power input, varying inversely as the latter at low levels of power input, and inversely as its cube root at high levels.

Further conclusions can be drawn from these relations with reference to development studies of relays and other electromagnets. The spring load and travel may be considered as fixed requirements, so far as the magnet design is concerned. The available power may be fixed by circuit considerations, or it may be related to the design by a requirement that the winding dissipate this power in the holding interval, a condition that imposes a minimum size on the coil and the magnet structure. Subject to this and some other limitations, there is a design choice of the dimensions and configuration which determine the magnetic circuit constants, the mass of the armature, and the eddy current conductance.

The preceding discussion has shown that, with an optimum choice of pole face area, the magnetic characteristics affect the time only with respect to the leakage factor, the ratio of leakage to useful flux. This factor may be reduced by using a "tight" magnetic circuit, but if this is done the factor tends to vary directly as the length of the magnetic path and inversely as the separation of the core and return members in relation to their cross sections. The leakage may be minimized by using a square outline for the magnetic path. The optimum speed magnet then has a specific configuration in which all dimensions are fixed in relation to the cross section of the magnetic members. This dimension then determines the mass of the armature.

For this optimum configuration, the power, the spring load, the mass, and the travel determine a level of pull which minimizes the operate time. This pull requires a certain armature cross section if saturation is to be avoided, as the effective pole face area is fixed in relation to the cross section. If the resulting armature mass is minor compared with the mass of the load, the operate time attainable varies with the cube root of the applied power, as in the cases discussed above.

With increased power, however, the optimum pull and armature cross section increase, and must eventually reach a level where the armature mass becomes the dominant portion of the mass term. As this condition is approached, any increase in power is offset by an increase in armature mass, such that a lower limit is imposed on the operate time proportional to the travel, corresponding to an upper limit to the average armature

velocity attainable with a neutral electromagnet. This upper limit is of the order of 100 cm/sec. Thus, for example, an operate time less than 0.25 millisecc cannot be attained with an armature travel of 10 mil-in, no matter how large the steady state power applied.

7 RELEASE WAITING TIME

Like operation, release is made up of an initial stage of flux change with the armature at rest, followed by a stage of armature motion, and the total time is the sum of the waiting time and the motion time. In release, the waiting time is usually larger than the motion time.

There are three distinct circuit conditions under which release occurs. These are:

Normal, or unprotected release, in which the coil circuit is open, and the only magnetomotive force maintaining the field is that of the eddy currents.

Protected release, in which the coil circuit is closed through a protective shunt, usually comprising a condenser and a resistance in series. The magnetomotive force comprises that of the eddy currents and that of the coil circuit transient.

Slow release, in which a sleeve or short circuited winding is used to maintain the field and delay release. The magnetomotive force is predominantly that of the sleeve or winding current.

The slow release case is the only one of the three for which the flux decay relation is accurately represented by equation (2). In protected release, the coil current transient is controlled by the condenser, as discussed below. In normal release, the variation in the field linked by different eddy current paths results in the changes in the field pattern discussed in Section 1. As noted there, equation (2) applies to this case only as a very crude first approximation.

In all three cases, the relation between φ and $\mathcal{F}$ applying is that for decreasing magnetization, as illustrated by the curve for $x = 0$ in Fig. 7. Unlike the linear relation applying in operate, corresponding to the constant reluctance given by (4), the decreasing magnetization curve has a hyperbolic character, and is asymptotic to the saturation flux φ'' . Residual magnetism results in a residual flux φ_0 , the intercept of the decreasing magnetization curve on the φ axis. An analytical treatment of the decreasing magnetization curve is given in a companion article,⁵ where it is shown to provide a satisfactory basis for predicting the release time of slow release relays, the third case above. The other two cases, to which the following discussion is confined, involve both non-linear mag-

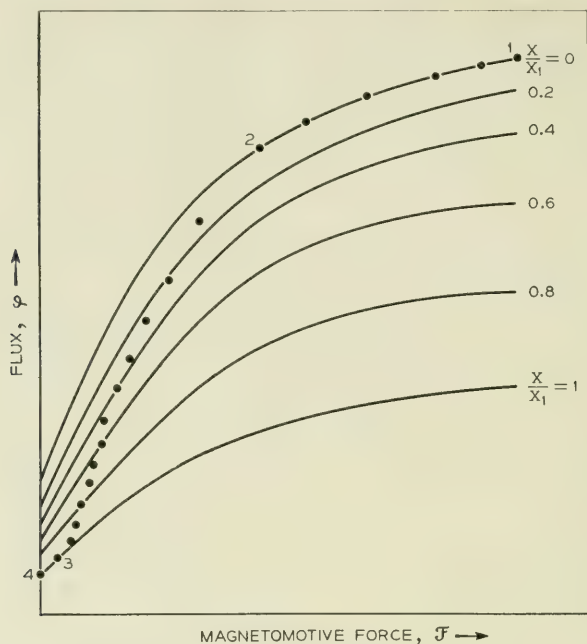


Fig. 7 — Field energy relations in release.

netization and a variable magnetic field pattern. In an approximation neglecting the latter effect, no advantage is derived from an accurate formulation of the former, and a linear approximation to the demagnetization curve may be employed.

NORMAL UNPROTECTED RELEASE

As indicated above, a rough approximation to the decay of the gap flux, and hence of the pull, may be obtained by assuming that this flux decays in conformity with equation (2), with $G = G_E$, where the eddy current conductance G_E has the same value as in operate. As a linear approximation to the demagnetization curve, the reluctance $\mathcal{R}$ can be taken as the closed gap reluctance $\mathcal{R}(0)$. Allowance may be made for residual magnetism by postulating a steady state magnetomotive force $\mathcal{F}_0 = \mathcal{R}(0)\varphi_0$. Then integration of (2) gives the following expression for the time t at which the flux has decayed to the value φ from its initial value φ_1 :

$$t = t_E \ln \frac{\varphi_1 - \varphi_0}{\varphi - \varphi_0}, \quad (25)$$

where $t_E = 4\pi G_E' \mathfrak{R}(0)$. Unless φ'/φ_0 is near unity, φ_0 may be omitted from this expression. If this is done, the expression may be re-written in terms of pull F in the form:

$$t = \frac{t_E}{2} \ln \frac{F_1}{F}, \quad (26)$$

where F is the pull at time t , and F_1 the initial pull. This expression follows from (25), for $\varphi_0 = 0$, because of the proportionality between F and φ^2 . If F is the spring load or tension for the operated position, and F_1 the steady state pull there, (26) is an approximate expression for the release time.

PROTECTED RELEASE

The commonly employed method of contact protection is to use a condenser-resistance shunt connected either across the coil, as in Fig. 8, or across the contact. Except for the steady state voltage of the condenser, the circuit relations are the same for the two cases. As a first approximation, eddy currents may be represented by the current in a short circuited secondary, as in Fig. 8. The effect of such a secondary is determined by the value of t_E representing the ratio of its inductance to its resistance, as this determines the ratio of its contribution to the flux to the time rate of change of the total field.

If perfect coupling and linear magnetization are assumed to apply, the circuit equations for Fig. 8 can be solved and expressions obtained for the flux, current and condenser voltage as functions of time after the opening of the contact. These expressions are similar to those for the simple C - R - L circuit without a secondary, except that they involve the time constant t_E as well as CR and L/R . The discharge may be either exponential or a damped oscillation, but while the discharge of the simple C - R - L circuit is always oscillatory for small values of C , the discharge in the circuit of Fig. 8 is only oscillatory for an intermediate

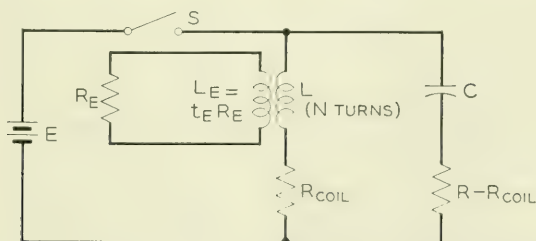


Fig. 8 — Coil circuit with capacitive contact protection.

range of capacity values. For large values of C , the discharge is damped by the coil resistance, and for small values of C it is damped by the currents in the secondary.

The latter case is essentially the condition applying for the small capacity values used for contact protection. Initially, the flux decay is opposed by the magnetomotive force resulting from both the eddy currents and the coil current. The latter discharges the condenser and develops a charge of opposite polarity. The initial rate of flux decay is therefore similar to that for exponential decay with a conductance equal to $G_E + N^2/R$, where R includes the external resistor as well as the coil resistance. The effect of the condenser charge is equivalent to a continuously increasing coil resistance, so that the decay approaches the rate that would apply for exponential decay with $G = G_E$, and the decay rate is then further increased by the reverse magnetomotive force resulting from the coil current reversal in the subsequent discharge of the condenser.

If CR is less than t_E , the effect of the coil circuit is to change the character of the flux versus time relation, without materially changing the time scale of the discharge. The initial decay is delayed, while the later decay is accelerated. Hence the release time is increased for a heavy load, corresponding to a high value of φ for release, while it may be slightly decreased for a light load, corresponding to a small value of φ . The utility of the coil-condenser circuit for contact protection results from the low initial rate of flux decay, which holds the induced voltage across the contact to a relatively low value in the initial stage of contact opening, when the contact separation is small.

Except initially, the predominant magnetomotive force is that of the eddy currents, and this results in a changing distribution of field intensity, as in the case of simple release. Hence the analysis of Fig. 8, as described above, is not applicable quantitatively, and is therefore not given here in detail. Qualitatively, the relations are similar, the protected release has a flux time relation similar in time scale to that for simple release, with a lower initial rate of decay, and a higher rate for the later stage. An illustration of this effect is included in the article by M. A. Logan⁴ cited above. No analytical treatment is available for determining the differences in release time between simple and protected release, at least for values of CR of the same order at t_E , as in the protection networks commonly employed.

8 RELEASE MOTION

It was stated in Section 7 that the waiting time in release is usually larger than the motion time. This, however, is not necessarily or in-

variably the case, and interest attaches to the conditions under which the motion time may become relatively large. Another important aspect of the release motion is the final velocity attained when the armature strikes the backstop. This impact results in a rebound which may, in the relay case, result in re-actuation of the contacts: rebound chatter. Rebound may be reduced by appropriate mechanical design, as discussed in an article by E. E. Sumner,⁸ but its amplitude is always, other things being equal, proportional to the kinetic energy of the armature in impact on the backstop.

Fig. 9 shows the force travel relations in release, corresponding to the operate relations of Fig. 4. The solid line, as in Fig. 4, represents the spring load, which has an operated value F_0 at the point marked 2. This corresponds to the similarly marked point in Fig. 7, which shows the corresponding φ versus $\mathcal{F}$ relations. The field decays along the demagnetization curve for $x = 0$ to the point 2, where load and pull are in equilibrium. Further decay results in a pull less than the spring load, and hence in a net accelerating force producing armature motion. The flux continues to decrease as indicated in Fig. 7, following the path 2-3-4. The pull follows the similarly marked curve in Fig. 9, related to the flux-travel path by equation (5).

The net accelerating energy is therefore represented by the area marked T , lying between the pull curve and the spring load. Thus only this portion of the energy V stored in the spring load appears as kinetic energy of the armature. The remainder of V is represented by the area

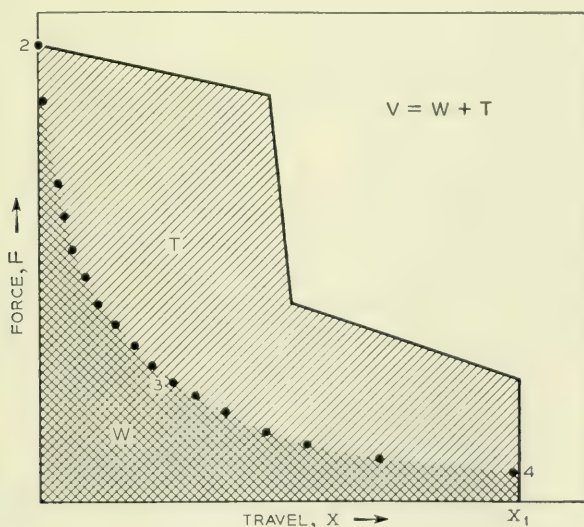


Fig. 9 — Load and pull relations in release.

marked W , lying below the release pull curve. This energy may be termed the magnetic drag: it is restored to the magnetic field and dissipated in the eddy current paths or in other circuits linking the field. Experimental studies have shown that the kinetic energy T of the armature in backstop impact may vary from 20 per cent to 90 per cent of V . Such a wide variation in armature energy has a proportional effect on rebound amplitude. Within certain limitations, the following analysis establishes the relations controlling the ratio W/V , and also serves for the evaluation of the release motion time.

EQUATIONS APPLYING

As the motion starts after an initial period of field decay, the relations applying are given approximately by (2) and (3), with $\mathcal{R}$ and F given by (4A) and (5A), and $G = G_E$. The use of (4A) and (5A) is justified by the fact that the lower portion of the demagnetization curve, which applies to the later stages of flux decay, is approximately linear. The assumption that (2) applies, with $G = G_E$, is justified by the fact, discussed in Section 1, that the initial rapid decay of the field in the outer layers of the core is followed by an exponential decay of the greater part (about 70 per cent) of the initial field. Thus if t_E is written for $4\pi G_E' \mathcal{R}(0)$ in (2), this expression applies in the latter stages of flux decay when the armature is at rest at $x = 0$. In this case, the value of t_E is a constant, even if the values of G_E and $\mathcal{R}(0)$ applying differ somewhat from those applying in the operate case.

Taking these equations to apply, it is desired to determine the solution for the initial conditions applying to the release case. This cannot be obtained by the approximations used for the operate case, as the simplifying condition of a nearly constant reluctance does not apply: the change of reluctance with x is a maximum for $x = 0$. The procedure that has been employed is the use of the analogue computer described in articles by A. A. Currie⁶ and E. Lakatos⁷ to obtain solutions for a restricted category of cases. The equations may be reduced to a form adapted to such solution as follows:

An expression for $\mathcal{R}/\mathcal{R}(0)$ obtained from (4A) is substituted in (2). Writing t_E for $4\pi G_E/\mathcal{R}(0)$, there is obtained:

$$\frac{C_L(1+u)}{C_L u} \varphi^2 + \frac{t_E}{2} \frac{d\varphi^2}{dt} = 0,$$

where u is written for x/x_0 , or $x/(A\mathcal{R}_0)$. By substituting in (3) the expression for F given by (5A) there is obtained an expression for φ^2 .

On substituting this in the preceding equation, there is obtained:

$$(1 + u)(C_L + u) \left(\frac{dV}{dx} - m \frac{d^2x}{dt^2} \right) + \frac{t_E}{2C_L} \frac{d}{dt} \left((C_L + u)^2 \left(\frac{dV}{dx} - m \frac{d^2x}{dt^2} \right) \right) = 0.$$

To reduce this expression to dimensionless form, the following substitutions are made:

f is written for $\frac{dV}{dx} / F_0$, where F_0 is the operated load, so that $f = 1$ at the start of the release motion.

τ is written for $2C_L t / t_E$, expressing the time t as a multiple of $t_E / (2C_L)$.

t_M^2 is written for $2mA\alpha_0 / F_0$. Thus t_M is the time for travel of the mass through the distance $A\alpha_0$ for a constant accelerating force F_0 .

K is written for $2C_L t_M^2 / t_E^2$.

The preceding equation then becomes:

$$(1 + u)(C_L + u) \left(f - K \frac{d^2u}{d\tau^2} \right) + \frac{d}{d\tau} \left((C_L + u)^2 \left(f - K \frac{d^2u}{d\tau^2} \right) \right) = 0. \quad (27)$$

This is the form of the release motion equation to which analogue computer solutions were obtained. It is a third order differential equation giving the travel, expressed as a multiple u of $A\alpha_0$, as a function of the time expressed as a multiple τ of $t_E / (2C_L)$. The boundary conditions correspond to zero initial values of travel, velocity, and acceleration, so that for $\tau = 0$, $u = du/d\tau = d^2u/d\tau^2 = 0$.

ANALOGUE COMPUTER SOLUTION

The analogue computer gives solutions to specific cases, and a specific case of (27) is defined by the values of C_L and t_M / t_E applying, and by the form of the relation between f and u , defining the shape of the spring load. To confine the cases considered to a manageable category, they were limited to those for which $f = 1$ for all values of u : the case of a constant spring load.

The constant C_L , the ratio $(\alpha_0 + \alpha_L) / \alpha_0$, is an inverse measure of the relative magnitude of the leakage field. As noted in Section 2, its value is usually in excess of 4. Solutions were obtained for two cases: $C_L = 3$ and $C_L = 5$. For $f = 1$, and a fixed value of C_L , the remaining param-

eter defining a specific case is t_M/t_E , or the corresponding value of K ($2C_L t_M^2/t_E^2$). For the two selected values of C_L , solutions were obtained for the following values of K : 0.05, 0.1, 0.3, 1.0 and 10.

The computer set-up for this problem is indicated in Fig. 9 of the article by E. Lakatos cited above.⁷ Solutions were obtained to (27) for $f = 1$ and the values of C_L and K cited above, for the given boundary conditions. The solutions were obtained in the form of machine drawn curves giving, u versus τ and $du/d\tau$ versus τ , over the range $0 < u < 3$. In accordance with the character of the net accelerating force, all these curves were smooth and monotonic.

For the conditions covered by the solution, these results give the motion time for a given travel directly in the form of the relation between τ and u . It is convenient to express the values of τ in terms of the corresponding values of t/t_M . The results for $C_L = 5$ are shown in Fig. 10 as curves of t/t_M versus u for various values of t_M/t_E .

The other result of interest is the ratio of the magnetic drag W to the spring load energy V , or of $(V - T)/V$, where T is the kinetic energy at the end of a given travel: the total travel in an actual case. T equals $m(dx/dt)^2/2$, or $2m(C_L A \mathcal{R}_0 \cdot du/d\tau)^2/t_E^2$. For $f = 1$, V equals $F_0 x$, or $F_0 u A \mathcal{R}_0$. It follows that T/V is given by:

$$\frac{T}{V} = \frac{mA\mathcal{R}_0}{2F_0 u} \left(\frac{2C_L}{t_E} \right)^2 \left(\frac{du}{d\tau} \right)^2,$$

and therefore by:

$$\frac{T}{V} = \frac{K}{2u} \left(\frac{du}{d\tau} \right)^2. \quad (28)$$

Corresponding values of u and $du/d\tau$ were read from the computer results and substituted in (28) to determine T/V and thus W/V . The resulting values of W/V for the case $C_L = 5$ are shown in Fig. 11 plotted against t_M/t_E for various values of u .

DISCUSSION OF COMPUTER RESULTS

The significance of the results shown in Figs. 10 and 11 can be more readily grasped by reference to representative values of the parameters involved. For relay electromagnets, representative values of $\mathcal{R}_0$ and A are 0.04 cm^{-1} and 1 cm^2 , respectively, for which $A\mathcal{R}_0 = 0.04 \text{ cm}$, or 16 mil-in. This distance is, for this case, the travel for which $u = 1$. For this value of $A\mathcal{R}_0$, an effective mass m of 10 gm, and an operated load F_0 of 2×10^5 dynes (200 gm wt), $t_M = 2$ millisecc. If the load were

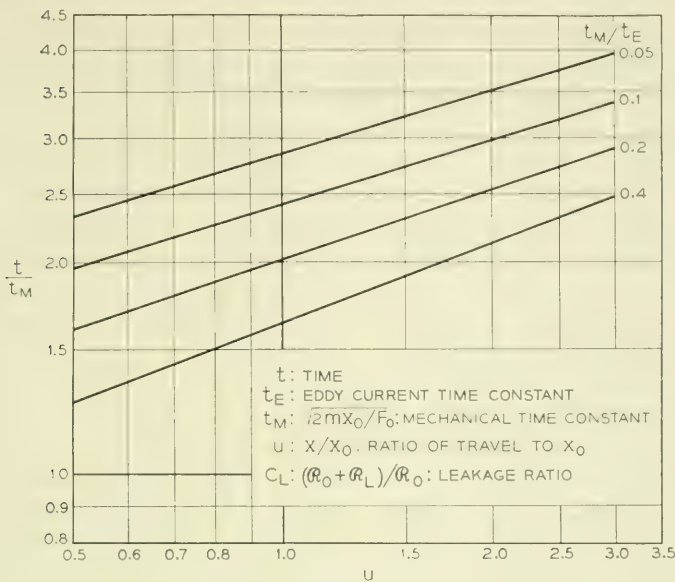


Fig. 10 — Release motion time relations.

doubled and the mass halved, t_M would be halved. These values of t_M are of the same order but smaller than the values of t_E generally applying, corresponding to the range in t_M/t_E covered by the figures. Similarly, the total travel usually lies in the range covered by the values of u , corresponding to travels up to 48 mils for the above value of $A\dot{R}_0$.

The curves of Fig. 11 show that the fraction W/V of the spring load energy absorbed by magnetic drag varies from 15 to 80 per cent over the range covered, and consequently that the kinetic energy at the end of the stroke, which determines the rebound amplitude, varies from 85 per cent to 20 per cent of V , or by a factor of 4 to 1.

The way in which W/V varies with u and with t_M/t_E is readily understood from physical considerations. Referring to Fig. 9, it is apparent that for a given time rate of field decay, the faster the rate of armature motion, the higher is the restraining pull and the less the kinetic energy T . For a short travel, or low value of u , the restraining pull will be larger than for a high value of u ; hence W/V decreases as u is increased. As t_M measures the time for a travel of $A\dot{R}_0$, or $u = 1$, for a force F_0 , t_M/t_E is a measure of the motion time relative to the time of field decay. Hence a large value of t_M/t_E corresponds to a condition where the motion is slow relative to the field decay, giving little restraining pull. Thus W/V decreases as t_M/t_E increases.

The effect of the magnetic drag on the motion time is apparent in the curves of Fig. 10. The motion time for a given value of u increases as t_M/t_E decreases, corresponding to increased drag. Thus a reduction in rebound by increased magnetic drag is, of course, accompanied by a longer motion time. It is of interest to note that the increasing acceleration results in an armature displacement which varies as the cube of the time, as in the relation applying in operate.

The results shown are for the case $C_L = 5$. For $C_L = 3$, corresponding to higher leakage, the results are similar, but the drag ratio W/V values are from 5 to 10 per cent lower than those for $C_L = 5$, and the values of t/t_M are correspondingly smaller. Thus the value of C_L applying has only a secondary effect on the release motion.

While these results are formally limited to the special case of a constant retractile force, the conclusions are of broader application. A load curve that decreases as the gap opens modifies the solution for the same operated load F_0 only in reducing the net accelerating force, increasing the drag ratio and the motion time. The residual magnetism, neglected in this discussion, has a similar, and minor, effect. The results show C_L to have little effect on the motion, which is governed primarily by the values of u and t_M/t_E applying.

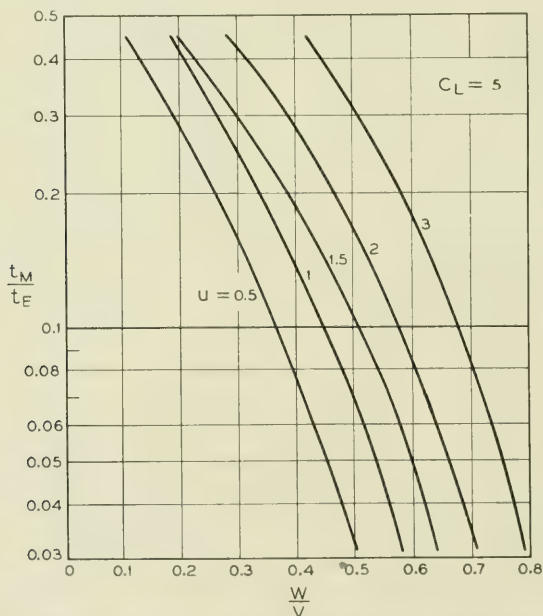


Fig. 11 - Energy absorbed by magnetic drag in release motion.

As minimizing rebound is of major importance in relay design, the relations controlling the kinetic energy to which it is proportional are of design interest. Most of the quantities appearing in the above relations, such as the spring load and the travel, are fixed by operating requirements or other design considerations. There is, however, considerable freedom in the design relations fixing the effective mass of the armature, particularly in selecting its axis of rotation. The preceding relations show that the magnetic drag ratio can be increased, and rebound reduced, by the decrease in t_M resulting from a decrease in effective mass. While such a decrease increases the ratio t/t_M , it results in a net decrease in the value of t , the motion time. A low effective armature mass is therefore advantageous both for fast release and for a high drag ratio for the reduction of rebound.

REFERENCES

1. R. L. Peek, Jr., and H. N. Wagar, Magnetic Design of Relays, page 23 of this issue.
2. R. L. Peek, Jr., Analysis of Measured Magnetization and Pull Characteristics, page 78 of this issue.
3. B. Wwedensky, Ann. d. Phys., **64**, p. 609, 1921.
4. M. A. Logan, Dynamic Measurements of Electromagnetic Devices, B. S. T. J., p. 1413, Nov., 1953.
5. R. L. Peek, Jr., Principles of Slow Release Relay Design, page 187 of this issue.
6. A. A. Currie, The General Purpose Analog Computer, Bell Lab. Record, **29**, pp. 101-108, Mar., 1951.
7. E. Lakatos, Problem Solving with the Analog Computer, Bell Lab. Record, **29**, pp. 109-114, Mar., 1951.
8. E. E. Sumner, Relay Armature Rebound Analysis, B. S. T. J., **31**, pp. 172-200, Jan., 1952.

Estimation and Control of the Operate Time of Relays

Part II—Design of Optimum Windings

By M. A. LOGAN

(Manuscript received September 4, 1953)

For each relay structure, there exists a best winding, once the circuit power has been specified. The best winding is that one which operates the relay in the least time. Methods for determining the best winding are developed.

On the basis of the operating time, the relay behavior is classed as mass or load controlled. The design method chosen depends upon this classification.

The design of windings for series connected relays is based on a method of determining equivalent single relays, the behavior of each corresponding to one of the series relays. This method is generalized to allow for different magnetic structures for the several relays. Each relay winding is then designed in turn, using the design data for its own type of structure.

The best winding design is not given directly by an explicit formula. Rather, methods are developed for determining the operate time for any winding. Then by choosing a range of windings, the best one is selected by interpolation.

INTRODUCTION

The selection of a relay for a circuit application, particularly so in common control systems, involves in part a determination of its operating and releasing time. In Part I of this article expressions are derived for the relations between these times, the design parameters of a relay, and the conditions of operation. These expressions are approximate, and have been developed with primary reference to the selection of favorable characteristics in design. The timing estimates they provide are of sufficient accuracy for the comparison of design alternatives. Once a basic design has been selected, the choice of windings and the prediction of the limiting times occurring in specific applications requires

a more detailed and exact procedure than can be obtained entirely from the application of these approximate relations. The information furthermore has to be quite versatile, to permit all new type variations to be included, such as winding, number of contacts and armature travel. It must, moreover, be in a form that permits determination of the upper and lower limits to the time observed in each specific type of relay as actually used, as in applications, interest attaches not only to the nominal conditions, but to all variations that may arise in actual manufacture and use.

Relay operation is complex, and in principle the nonlinear differential equations which describe it can be solved exactly only by computers. Even if this is done, the results are subject to any uncertainty that exists as to the exact form of the relations that apply. The representation of non-linear magnetic material properties, the discontinuous load-travel characteristics, the eddy current effects, etc., do not make such an approach attractive.

The relay, however, is a perfect analog of itself. With the magnetic structure set, controlled models can be built to include dimensional and magnetic material variations. With these, exact solutions to a variety of conditions can be determined. With these data available, approximate solutions can be used for interpolation and extrapolation, determining the effect of small variations from the tested conditions.

This part of the article will exhibit the form of data presentation in two classes, mass and load controlled operation. It includes the theory used in selection of the forms, and the correction methods used for estimation of variations from the standards. The initial part will be concerned with a single relay operating in a local circuit. The latter part will consider series and series-parallel operation of similar structures but not necessarily identical windings. This latter problem is solved by determination of an equivalent relay in a local circuit for each of the several relays. The earlier analysis then can be applied to each in turn.

The order of analysis can be reversed. That is, given required operating times, windings can be determined which will provide these times.

In this article, a best winding is (1) that winding which, for the specified applied power, results in the minimum operate time or (2) that winding which, for a specified operating time, requires the minimum power. A unique solution exists.

Existence of Best Winding

Fig. 1 shows measured operate time of a relay for two different de power conditions, versus number of turns in the winding. That a best

winding, in the definition of this article exists, as exhibited by the minima on Fig. 1, was shown in Part I to be a consequence of equation (12). This best winding and its determination is the basic subject of this article.

OPERATE TIME — SINGLE RELAYS

As in Part I, the operating time of a relay is considered as made up of three stages: (1) the waiting time while the armature remains at the backstop and the pull builds up to equality with the back tension, (2) the motion time beginning at the end of the waiting time and continuing while the armature moves from the backstop to the position of the earliest contact, and (3) the stagger time during which the armature actuates all of the remaining contacts.

The mass controlled case, as a practical matter, simplifies to a determination of only the first two without regard for the spring load, with the displacement of the armature taken to the latest contact in the array. The armature pull builds up to values in excess of the load during its early motion and the velocity is so high during the stagger time that this small interval can be included as part of the motion time. This treatment is essentially that of the three stage approximation of Part I.

The load controlled case simplifies to a determination of the time required for the pull to build up to the maximum load, including the back tension, with most of the relatively slow motion taking place during this pull buildup. The remainder of the motion time is accounted for as an empirically determined correction factor in the expression for time of pull buildup. This treatment corresponds to the single stage approximation of Part I, with a correction term to account for the additional motion time.

Fig. 1 shows graphically the basic characteristics of the two types of operation. This chart displays the operating time of the same relay under the conditions of 1.0 or 5.7 watts, and three contact load conditions: 12, 18, and 24 contact pairs. For each of the six conditions, a curve is drawn showing the effect of the number of winding turns. It is clear that (1) there is a best number of turns for each case, (2) for the 5.7 watts the best number of turns is the same for all three loads, (3) for the 5.7 watts case the difference in operating time is only 0.1 millisecon in 5.6 millisecon, between 12 and 24 contact pairs, (4) for the 1.0 watt case the best number of turns increases with the number of contact pairs, and (5) for the 1.0 watt case the operate time is double for 24 compared to 12 contact pairs. These strikingly different behaviors form the basis for division of relays

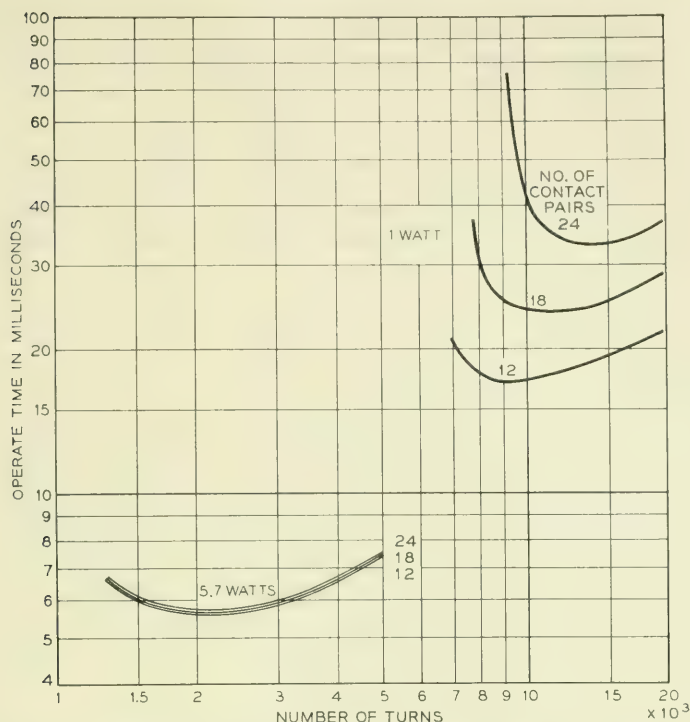


Fig. 1 — Chart showing typical mass and load controlled operation.

into two classes called mass and load controlled, and furthermore empirically establish which of the three stages of operate time predominates. This distinction corresponds to that made in Part I in the discussion of equation (12). For the lower power cases the term involving the spring load predominates, for the higher power cases the mass term predominates. As this equation shows, there is necessarily a transition range where the operation time merges from one class into the other. Both types of charts are extended to include this range. A rule of thumb for an estimate of whether a time under discussion is in the mass or load controlled case is that if the time is less than

$$t_M = \sqrt{\frac{27m(x_1 - x_3)}{F_3}} \quad (1)$$

it is mass controlled; otherwise it is load controlled. The derivation of this bound will be discussed when mass controlled operation is considered.

The single relay design data are general enough to include resistors

external to the relay winding. Wherever a resistor or power term appears, the sum of all resistances or powers is indicated.

Maximum Versus Average Operate Time Presentation

In what follows, load controlled operation data are given in terms of maximum times, whereas mass controlled data are average. Either choice could be used for both. The data for converting to average or maximum respectively will be described. However, the choices made here are consistent with the normal use of the data.

Mass controlled, sometimes called speed, relays operate with relatively large power and ordinarily are used in common control circuits where many events occur in succession. The total number of similar common equipments necessary in an office is related to the control circuit holding time. For a system design, the cost of power is balanced against the cost of equipments to arrive at a minimum office cost.¹ Now the maximum holding time per call is never the sum of the maximum possible times of each of the several relays operating in succession. It rather is more nearly the sum of the averages, increased by considerably less than the common maximum to average ratio of one of them. For example, if n relays are assumed to have a distribution approaching normal, an analytical expression for the probable maximum is:

$$\delta t_{av} + \left(1 + \frac{\frac{t_{max}}{t_{av}} - 1}{\sqrt{n}} \right) \sum t_{av}, \quad (2)$$

where δt_{av} is an allowance for short time deviations of the manufactured product from the long time average. Thus for mass controlled relays the most directly applicable type of data is in the form of average time, and maximum to average time ratio.

For load controlled relays just the opposite is true. Here the operating times are relatively long, either because the power drain is to be kept to a minimum, as in a long holding time circuit, or it is an event which takes place while several successive mass controlled events occur. For either case, it is a single event and only its maximum duration is desired. Here then the most directly applicable type of data is maximum. In addition to these, other data for minimum times are needed for studies of timing when two parallel circuit paths occur.

Because how far the relay armature has to move is the outstanding variable in mass controlled relay timing, these charts are prepared for each nominal distance which can be chosen. When there is no concern as

to relative actuation time of the several contacts on one relay, the smallest armature travel is ordinarily provided. When a definite sequence is needed for circuit reasons, the actuating means is arranged to guarantee operation of certain contacts before the others. This necessitates the provision of a greater armature motion.

The principles used in preparing these other types of charts are identical to those used for the two types which are developed specifically in this article.

Local Circuit Load Controlled Operation

Power Given

The best winding for a load controlled relay is not here given explicitly by a formula, but rather is found indirectly by developing a method for determining the operate time for any winding. Then by assuming a range of these, the best one can be selected by interpolation of these derived data. In Part I, the waiting line for a linear system was shown as equation (9) to be:

$$t_1 = \frac{4\pi}{\mathfrak{R}_3} (G_C + G_E + G_S) \ln \frac{1}{1 - I_0/I}, \quad (3)$$

with I_0/I substituted for ν . Without motion time, saturation, eddy currents, or non-linear effects, Part I of this article also shows that the best winding is that for which the turns are selected so that $NI_0/NI = 0.715$.

Taking the other variables explicitly into account complicates the determination and is at best only approximate. Instead they are included through this interpolation approach. In a previous article² a better representation for the core eddy current constant was shown to be:

$$G'_E = G_E e^{-G_E/(G_C + G_S)} = \text{effective core eddy current constant.}$$

Substituting this and explicitly indicating a correction term for the small motion time, the form best suited for the present discussion becomes:

$$t_0 = (1 + t_2/t_1) L_1 (G_C + G_S + G'_E) \ln \frac{1}{1 - q}. \quad (4)$$

Except for the correction term for the motion time, t_2/t_1 , this is still a linear equation with all factors known. Omitting the motion time correction, it can be solved either by numerical substitution or by a nomogram, once a value for L_1 has been chosen. A conservative value is the average one turn inductance, for the late contact critical load point where $x = x_3$. A more accurate value is determined through use of both

the inductance at the open gap and at the critical load gap. These are weighted in proportion to the ampere turns developed at each location. The ampere turns necessary to overcome the back tension is one factor, and the other is the difference between this first value, and the total required for operation at the critical load gap. This weighting yields an effective inductance value intermediate between the two extremes for each problem, but a new chart does not have to be prepared. An operate time using the chart is first determined. This time is then adjusted by the ratio of the effective inductance applying and the inductance used for making up the chart. This is exact, as the inductance term appears only as a direct multiplier.

A typical nomogram is shown in Fig. 2. It already includes the motion time correction, whose determination will be discussed. The dashed line

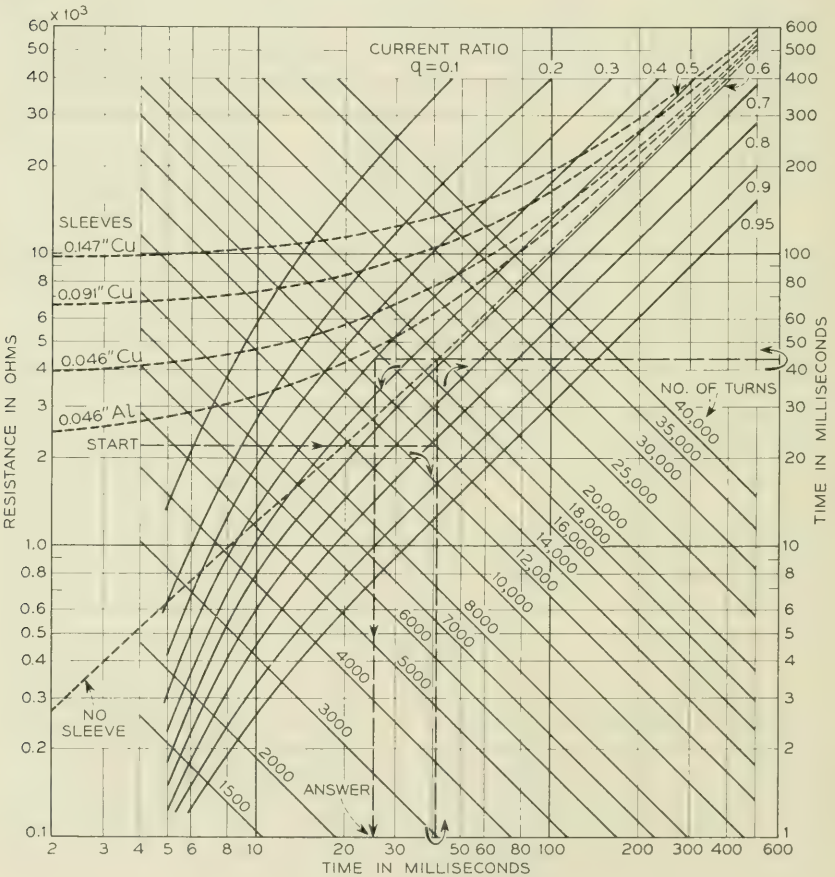


Fig. 2. — Nomogram for solution of load controlled operate time.

shows the successive steps taken in determining the numerical value for a specific case.

The dc circuit resistance is determined by the known circuit voltage and specified power. Entering the chart on the left ordinate at this resistance and proceeding horizontally to the right, intersections with increasing number of turns lines, sloping downward to the right, are found, and the appropriate one is chosen. Dropping vertically to the abscissa, the winding time constant $t_c = L_1 G_c$ is determined.

To this is next numerically added the known effective core eddy current and a sleeve (if any) time constant by returning vertically to an intersection with the appropriate core indicated as "no sleeve," or core plus sleeve, curve. Proceeding horizontally to the right hand ordinate scale from this intersection, the time constant multiplier of the $\ln$ term in equation (4) is determined.

The multiplication of these two factors is accomplished by proceeding to the left along this same horizontal line, to an intersection with the proper q line, sloping downward to the left. Vertically below this last intersection is the operate time.

Initially, tentative q lines are drawn, omitting the motion time correction. These lines are straight with a positive 45° slope. Then with an actual relay whose just operate current has been measured, operate time measurements are made, keeping the final current, and hence q , fixed at several values in turn. This is done by adjusting an external resistance and battery voltage over a wide range, effectively changing N^2/R . These measured data are plotted, following the same steps through the nomogram, except the last intersection is with the measured time vertical, rather than the known q . This provides several empirically determined q lines. These are used as templates, to progressively alter the shapes of the tentative straight q lines drawn earlier and shift them to the right. This adjustment then introduces the motion time correction.

It has been found empirically that for large operating times, the motion time correction factor has a value of about 0.1. There is no definite division between mass and load controlled operation, but as the total time decreases, travel time becomes more important. The correction factor increases to a value of about 0.5, in the transition range. For completely mass controlled operation, there is no q effect, so the q lines must all eventually converge.

As stipulated above, this chart applies to the current, flux and pull range where the first approximation magnetic constants of the structure are applicable. For this reason, the tests for the motion time corrections are made under conditions meeting these restrictions.

These curves should be corrected for magnetic saturation if there is a

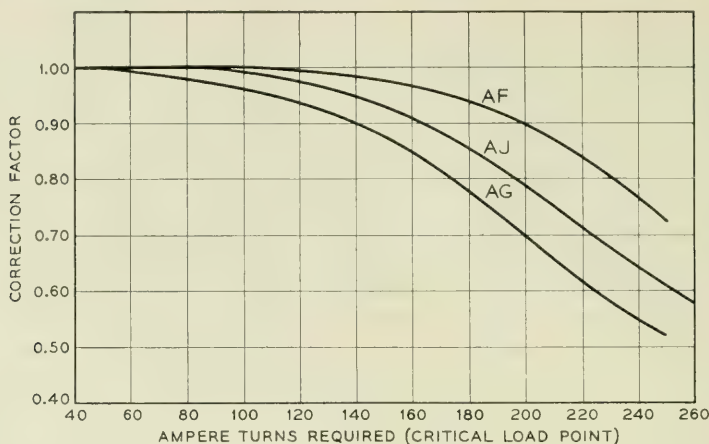


Fig. 3 — Correction of computed operate time for magnetic saturation.

large load at the critical load point and the ampere turns needed are in the saturation region for the core or armature. In the saturation region the current builds up faster than the linear time constants indicate, and, therefore, the indicated operate times are too large. Correction factors are determined by graphical integration of the integral form of the flux rise equation expressed as a ratio to the linear relationship:

$$\text{Saturation Correction Factor} = \frac{\int_0^{\varphi_s} \overline{NI} - \bar{N}i \, d\varphi}{-L_1 \ln(1 - NI_0/NI)} \quad (5)$$

These corrections are plotted as a function of the just operate ampere turns NI_0 , with the final NI as a parameter for each curve. Actually, because the correction factor is only of the order of 20 per cent maximum, assuming the final winding ampere turns are well into the saturation region, it is found that a single curve for any one type of relay fits all the computed points to an accuracy of a few per cent, and generally is used. Such composite curves are shown in Fig. 3 for the three types of wire spring relays.

A method has now been established for determination of the operate time of a relay with two restrictions (1) that the final ampere turns will operate the relay and (2) that the relay is in the load controlled class. An indication of the latter is whether the operate time determined is in the time region where the q curves are decreasing in curvature. For Fig. 2, 10 milliseconds is taken as the lower bound for load controlled relays.

Now the determination of an optimum winding for a particular

problem can be finished. Starting with the power given and choosing an arbitrary number of winding turns, the operate time is determined as described above. If it falls into the load controlled class, then a different number of turns is next assumed and a second time determined. This procedure is repeated until a time curve versus number of turns similar to Fig. 1 can be plotted, including a minimum. The best winding is this minimum. A different curve and optimum number of turns will apply to each contact load assumed. Three computed curves corresponding to the measured 1-watt curves are shown on Fig. 4. These were determined using effective inductance values described earlier.

Finite wire sizes permit only certain number of turns to be physically realizable when the resistance has been specified, without splicing two gauges of wire. The nearest gauge on the coarser wire size side is chosen, resulting in slightly too many turns. Note that the curves rise less steeply on the high turn side and the time penalty therefore is less than if too few turns were supplied.

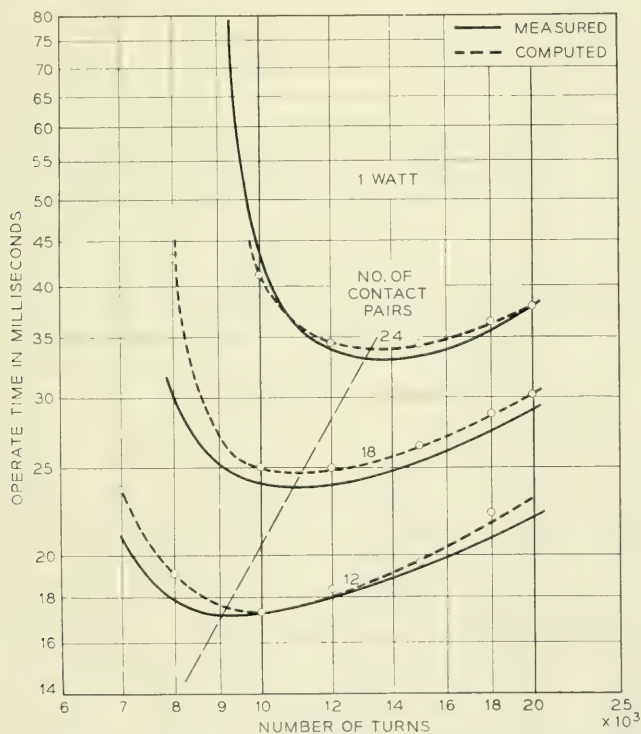


Fig. 4 - Typical computed maximum operate time curves for load controlled relays.

Maximum Time Given

For a specified maximum time, the above process is repeated for several assumed circuit powers until the specified time is bracketed. Then by interpolation of the optimum times indicated, the minimum power, maximum resistance and optimum turns are determined. As an example, Fig. 4 can be used to demonstrate the method. Assume that the three curves were computed for different circuit powers, rather than for different contact spring loads. A line is drawn through the minima. The intersection of this line with the required operate time determines the number of turns. For instance, if 30 millisecc were required, the turns would be 12,500. The circuit power at this same intersection can be interpolated for, using the known circuit powers associated with the three curves.

It is not economical to have a different winding for each spring combination. For this reason, a winding is designed for the maximum spring load and then used for smaller loads. The operate time will always be less with the smaller loads.

Measurements of Time Curves

Before considering mass controlled operation, the simulation of windings will be discussed. In the above description for establishing the q curves, it was pointed out that by adjusting an external series resistor and the battery voltage to maintain constant final current, the coil constant N^2/R was altered without changing q . This can be further extended to permit simulation of any winding for test purposes providing only that the experimental coil fills the winding volume as much or more than the coil to be simulated. For this purpose, a special test winding

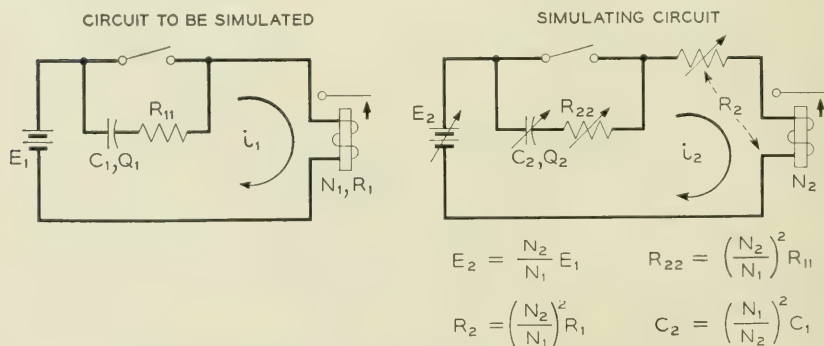


Fig. 5 — Simulation of winding circuit for timing tests.

always is used which completely fills the winding volume. Hence any winding which can be designed also can be simulated.

The conditions which must be fulfilled for perfect simulation derive from Lenz's Law. It is essentially an impedance transformation technique keeping the magnetic flux invariant, with the assumption that a winding can be considered as a lumped rather than a distributed network. This is equivalent to stating that at any instant the current flow is the same in every turn and there is no propagation time involved. This is true for times involved in electro-magnets.

Fig. 5 shows a circuit to be simulated, in which all the components with subscripts 1 have been given. The simulating circuit has only the number of turns N_2 , of the test winding, given. The other four elements must be determined. After switch closure, the exact differential equations applying are

$$\begin{aligned} N_1 \frac{d\varphi_1}{dt} &= E_1 - i_1 R_1, \\ t = 0; \quad i_1 &= i_2 = 0. \\ N_2 \frac{d\varphi_2}{dt} &= E_2 - i_2 R_2. \end{aligned} \quad (6)$$

Now for equality of magnetic flux, the two rates of flux change must be identical at all times including the first instant. Inserting the initial boundary conditions and equating the two rates, we have

$$\frac{E_1}{N_1} = \frac{E_2}{N_2}. \quad (7)$$

At infinite time, the same magnetomotive force must apply to both circuits for equality of final flux. Equating these, and cancelling the 4π factor,

$$N_1 I_1 = N_2 I_2. \quad (8)$$

Noting that

$$I_1 = \frac{E_1}{R_1}, \quad (9)$$

$$I_2 = \frac{E_2}{R_2},$$

we have, after using (7) and rearranging,

$$\frac{N_1^2}{R_1} = \frac{N_2^2}{R_2}, \quad (10)$$

which states that the coil constants must be equal. After steady state has been reached and the switches opened, the two differential equations are:

$$\begin{aligned} E_1 &= i_1(R_{11} + R_1) + N_1 \frac{d\varphi_1}{dt} + \frac{1}{C_1} \int_0^t i_1 dt, \\ E_2 &= i_2(R_{22} + R_2) + N_2 \frac{d\varphi_2}{dt} + \frac{1}{C_2} \int_0^t i_2 dt, \end{aligned} \quad (11)$$

$$\text{at } t = 0, \quad i_1 = \frac{E_1}{R_1}, \quad i_2 = \frac{E_2}{R_2}, \quad Q_1 = Q_2 = 0.$$

Multiplying the first equation by $N_1 C_1$ and the second by $N_2 C_2$, and equating term by term for equality at all times and equal magnetomotive forces, two additional equations result:

$$\begin{aligned} (R_{11} + R_1) C_1 &= (R_{22} + R_2) C_2, \\ N_1^2 C_1 &= N_2^2 C_2. \end{aligned} \quad (12)$$

The four defining equations, with some further rearrangements using (10), are shown in Fig. 5 as the relations applying for equality of magnetic flux in the two circuits. As the mechanical behavior of a magnetic structure is completely determined by the magnetic flux, it follows that the mechanical performances will be identical and timing measurements made with the simulating circuit will represent the actual circuit. This is true whether the relay is mass or load controlled. It includes eddy current effects and whether or not there is motion of the armature.

Local Circuit Mass Controlled Operation

Typical mass controlled operate time curves are shown in Fig. 6. These are for two different armature travels, indicated as short and intermediate. The curves are characterized by the circuit power used for each, with the coil constant plotted along the abscissa and the corresponding operate time as the ordinate. As mentioned earlier, these curves are substantially independent of contact spring load, and are plotted for average conditions, including an averaging of the time for the first and the last contact to be actuated.

It will be noted that the best coil constant is not independent of the circuit power, decreasing continuously as the power is increased. Also by increasing the circuit power from 2.3 to 23 watts, the operate time is decreased by a factor of a little less than 3, which is nearer the square

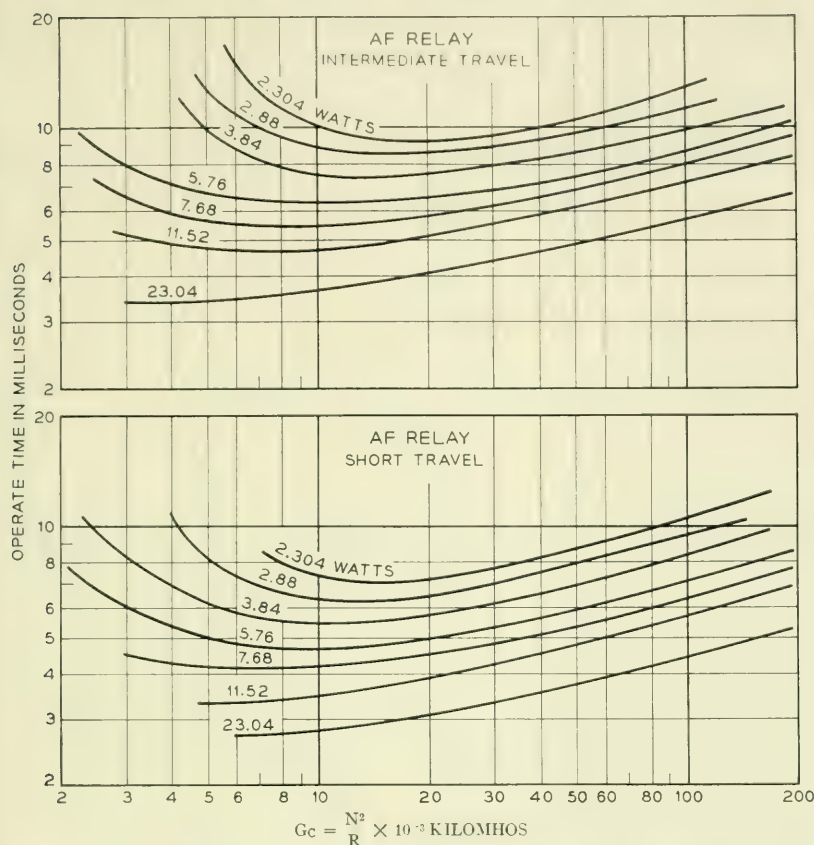


Fig. 6 — Mass controlled operate time tests.

root than the cube root of the power ratio, developed in Part I, for the mass controlled case. This is partly attributable to the fact that waiting time is included, and partly to the fact that the 2.3 watts case is in the transition range between mass and load controlled operation.

For the operating region where coil constants are larger than the best, the time curves are parallel and increasing. Considering any one vertical line representing a particular winding, increasing the power by increasing the battery voltage will always decrease the operate time. For curves plotted as in Fig. 1, where the battery voltage is kept constant and the time curves are plotted with winding turns as the independent variable, parallel curves again obtain. Thus an increase in circuit power obtained by keeping the voltage constant and reducing the circuit resistance always will reduce the operate time. Conversely, adding any series

impedance, not including a capacitance, will always increase the operate time. This will be made more evident when the operate times of relays with their windings in series are considered.

The curves can be used in either of two ways (1) given the relay circuit for which the applied power and coil constant can be computed, the average operate time can be found from the chart, or (2) given a required average operate time, the necessary circuit power and coil constant can be found from the chart.

These curves can be plotted in this form to exhibit the best winding directly because of the independence to contact spring load. For the load controlled case, the contact spring load was an essential parameter. For the mass controlled case, the armature travel becomes the outstanding parameter to be considered, but there are only two or three of these. All the other factors except contact load also enter and need to be evaluated for two reasons, (1) to provide an estimate of the range in operate time to be expected and (2) to adjust experimental measured time data to average. For any experimental setup, it is seldom possible to provide a structure which is average in every respect. For any one structure, however, all the factors known to affect its performance can be measured. Comparison of these to the manufacturing specifications locates the experimental setup in the universe of all relays as regards each of the factors. It is thus necessary to develop representations relating each of these factors to the operate time, which will suffice for the two uses named above. The development of these relationships will be the subject of the following sections.

Waiting Time

The waiting time, whether an electromagnet is mass or load controlled, is given by the same form of equation used for the total operate time of a load controlled relay:

$$t_1 = L_1(G_C + G'_E) \ln \frac{1}{1 - q_1} . \quad (13)$$

where now q_1 is determined by the armature back tension F_1 ; L_1 applies to the open gap; and no sleeve conductance is present. For the present purpose, it is desirable to rewrite this equation in terms of the fundamental parameters of the relay. For the open gap case the magnetic material is operated in its linear region and the open pole face gap provides additional linearity. For these reasons the expression is quite accurate. The sketches of Fig. 7 show the factors to be used. The value of

L_1 ,⁴ the inductance for one winding turn is:

$$L_1 = 4\pi \left(\frac{1}{\omega_L} + \frac{1}{\omega_0 + x_1/A} \right). \quad (14)$$

Also

$$NI = \sqrt{G_c W}. \quad (15)$$

From the network of Fig. 7,

$$\varphi_G = \frac{4\pi Ni}{\omega_0 + x_1/A}. \quad (16)$$

The armature force developed is

$$-F = \frac{\varphi_G^2}{8\pi A}. \quad (17)$$

When the magnetic pull has reached F_1 , the waiting time is over, determining φ_G , which in turn determines NI_0 , and hence q .

Substituting (14), (15), (16), and (17) in (13) we have

$$t_1 = 4\pi \left[\frac{1}{\omega_L} + \frac{1}{\omega_0 + x_1/A} \right] (G_c + G'_E) \ln \frac{1}{1 - (A\omega_0 + x_1) \sqrt{\frac{F_1}{2\pi A G_c W}}} \quad (18)$$

as the desired expression for the waiting time in terms of the fundamental constants. Use will be made of this later when an expression for motion time has been developed.

Motion Time

The motion time immediately follows the waiting time and is the time required for the armature to move from the backstop position x_1 , to the location of the last contact x_3 . From Fig. 1, it is permissible to omit any consideration of contact spring load, and consider the motion to be controlled entirely by the armature mass and magnetic pull developed after the waiting time is over. The determination of motion time is simplified because the initial velocity and net force on the armature are both zero, the latter following from the definition of the end of t_1 . In any problem involving motion, there must be established the initial velocity, position, and a suitable expression for the ensuing force. With

these factors, the differential equation of motion can be solved, and the time for a given travel determined.

Thus there is now needed an expression for the armature force developed by the pole face gap flux as a function of time. Briefly, the basis for the method to be described is the assumption developed in Part I, that the winding flux continues to rise during the initial motion interval in the same way it would have risen if the armature had not moved. With this assumption, an expression for the motion can easily be derived. With this expression it then can be shown that the armature does spend most of its time in the close vicinity of the backstop, justifying the initial assumption.

The present approach differs from the more general one of Part I. It uses the results there developed regarding the behavior of mass controlled operation, to simplify the initial motion equation and obtain an expression for motion time in a form better suited for the present purpose.

Armature Force Rise and the $\dot{F}$ Concept

As an introduction to the method which will be used, consider the general character of flux build-up in a relay. It will have a shape somewhat similar to an exponential curve. Now with the armature at the backstop, from the first approximation magnetic network of Fig. 7, a fixed portion of this winding flux will pass through the pole face gap, with the result that the pole face gap flux curve will also have the same general character. Such a measured curve is shown in Fig. 8, as well as the curve when the armature is free to move.

Now the armature force developed is proportional to the square of

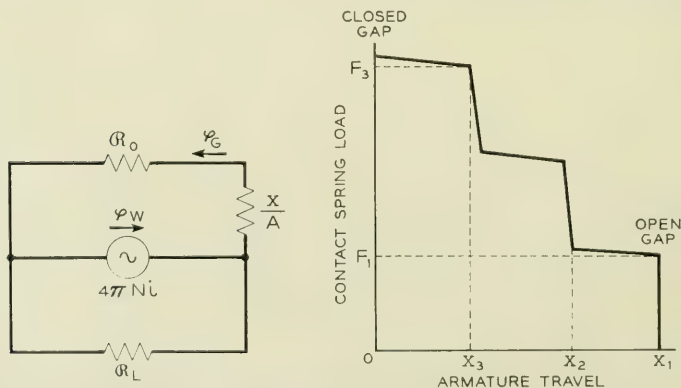


Fig. 7 -- Schematics of nomenclature applying to operate time.

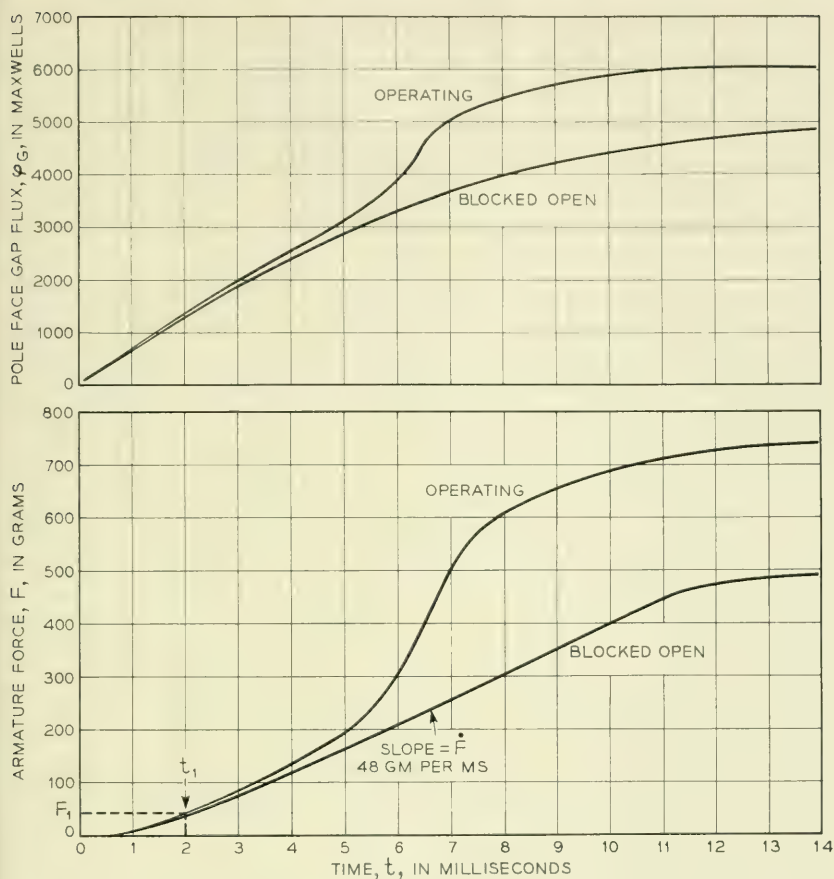


Fig. 8 — General character of winding flux and armature force rise.

the pole face gap flux as given by (17). Hence, the initial force rise will be parabolic, as shown by the lower portion of the curves of Fig. 8. After a long enough time however, the flux will have reached its maximum and the finally developed force will be constant with time. This is shown by the force curves approaching horizontal lines to the right in the same force diagram. As the force curve is continuous, there also must be an inflection point in the curve, and the entire curve has the general shape as shown.

The waiting time t_1 occurs while the magnetic force builds up to F_1 , as indicated, amounting to 2 millisecc for the example shown. This time generally includes all of the parabolic part of the curve. Following the waiting time, the initial force build-up necessarily is almost linear be-

cause of the inflection of the curve. Two other factors, caused by the ensuing motion, act to provide even more force than the open gap curve indicated by Fig. 8. The first is that a smaller gap takes a larger portion of the winding flux than assumed in the diagram, and the second is that the winding flux rises more rapidly and to a higher value than the open gap curve assumed. The operating force curve shown includes these effects. The relay operates in about 7 millisecc and for 6 of the 7 millisecc, the two force curves differ very little.

Hence, for a relatively long interval after the start of the motion time, a very simple force relationship holds, namely that the force is directly proportional to time, and the proportionality factor is the slope of the straight portion of the curve, designated as $\dot{F}$. For the example given, this amounts to about 48 grams per millisecond. By using Newton's Law, an expression for the motion time can be written.

$$m \frac{d^2(x_1 - x)}{dt^2} = \dot{F}t, \quad (19)$$

from which, with the initial velocity zero

$$m(x_1 - x) = \frac{\dot{F}t^3}{6}. \quad (20)$$

Rearranging, the expression for motion time becomes:

$$t_2 = \sqrt[3]{\frac{6m(x_1 - x_3)}{\dot{F}}}. \quad (21)$$

The factor $\dot{F}$ will now be considered.

Derivation of Expression for $\dot{F}$

Experimental—The factor $\dot{F}$ can be determined graphically. This brings in the second order effects such as quality of iron, cross section, residual magnetism, fit of parts, etc. For this purpose three working curves characteristic of the structure, all at the open gap, are first prepared for the particular winding:

- (a) The dynamic flux rise curve,²
- (b) The static flux curve versus ampere turns, and
- (c) The static pull curve versus ampere turns.

Choosing a time on curve (a), the corresponding instantaneous flux transferred to curve (b), determines the equivalent magnetomotive force. This transferred to curve (c) yields the instantaneous magnetic force. Repeating this for other times in the range of interest, an armature force

curve like Fig. 8 is established. The waiting time is read directly, corresponding to F_1 . The slope of the ensuing linear force range determines $\dot{F}_1$. This, with equation (21) completes the determination of mass controlled motion time. Of course, the final pull developed has to exceed the operated load. This check is made from another pull curve taken at the armature gap x_3 , using the known final ampere turns and the load.

Analytical - For our present purposes, an analytical relation, expressed in the fundamental constants, similar to that for the waiting time is needed. Its derivation follows:

The solution which will be developed is based on linear circuit theory. This necessarily implies exponential flux rise, which is not exactly true. However, the relation is dimensionally correct and accurate to better than first order. Then the use which will be made is to determine the motion time of an electromagnet as the parameters are varied one at a time. These are plotted as ratios to one of them, chosen as a reference. By this means the ratio curves become accurate to better than second order and provide excellent correction factors for actual measured data.

For a linear circuit, the pole face gap flux will increase, after the winding circuit is closed to a battery, with the same time constant as the winding:

$$\varphi_g = \frac{4\pi NI}{\mathfrak{R}_0 + \frac{x}{A}} (1 - e^{-t/T}), \quad (22)$$

where $T = L_1 (G_c + G'_g)$ and $I = E/R$. Then the pull:

$$-F = \frac{\varphi_g^2}{8\pi A} = \frac{1}{8\pi A} \frac{(4\pi NI)^2}{\left(\mathfrak{R}_0 + \frac{x}{A}\right)^2} (1 - e^{-t/T})^2, \quad (23)$$

$$\frac{dF}{dt} = \frac{4\pi(NI)^2}{AT\left(\mathfrak{R}_0 + \frac{x}{A}\right)^2} [e^{-t/T}(1 - e^{-t/T})]. \quad (24)$$

The bracket term is of the form $a(1 - a)$, $0 < a < 1$, which has a maximum value of 0.25, and from $0.2 < t/T < 1.5$ is between 0.15 and 0.25. This range corresponds to the maximum slope of the armature force versus time plot in Fig. 8. Arbitrarily 0.2 is chosen and the expression for the maximum rate of force rise becomes:

$$\dot{F} = \frac{4\pi(NI)^2}{5AT\left(\mathfrak{R}_0 + \frac{x}{A}\right)^2}. \quad (25)$$

Substituting for T and rearranging:

$$\dot{P} = \frac{(NI)^2}{5(G_C + G'_E)(A\mathcal{R}_0 + x_1) \left(1 + \frac{\mathcal{R}_0}{\mathcal{R}_L} + \frac{x}{A\mathcal{R}_L}\right)}. \quad (26)$$

Substituting (26) and (15) into (21), the final expression for motion time becomes:

$$t_2 = \sqrt[3]{\frac{30m}{W} (x_1 - x_3) \left(1 + \frac{G'_E}{G_C}\right) (A\mathcal{R}_0 + x_1) \left(1 + \frac{\mathcal{R}_0}{\mathcal{R}_L} + \frac{x_1}{A\mathcal{R}_L}\right)}. \quad (27)$$

This, with equation (18) for the waiting time, forms the basis for the variation effects which now will be determined. Its region of accuracy is for windings with turns exceeding the best, as it presupposes the relay will always operate. It thus is applicable in the region where the time curves are parallel. For studies of lower numbers of turns, the three stage approximation of Part I does not have this limitation.

Operate Time Variations

For a particular electromagnetic structure the factors in equations (18) and (27) can be measured. These measured factors can then be compared to the manufacturing specification and estimates made of average values for each. With these average values, the waiting time, the motion time and then the sum, which is the operate time, can be computed. This is the reference condition.

Each factor is then varied over its appropriate range, one at a time, and the computation repeated. The ratio of these variation times to the reference condition can then be plotted. Fig. 9 shows the chart for the wire spring type relay. For this chart, the range for each factor was taken as 3 without regard to the actual range. Similar charts have been prepared for other types of electromagnets, including quite a size difference. These ratio charts all agree remarkably well when plotted in this way. For this reason, for early estimates of any structure in the mass controlled class, this chart is entirely adequate for estimates of variations.

The basic assumption made in this method of determination is that for small variations, the interactions are negligible and a separated solution of products, one for each factor, is applicable. Checks made by varying two at a time confirm that essentially the presumption is fulfilled.

The G_C curve does not exhibit a minimum, as do the measured curves of Fig. 6. This results from the simplifying stipulation made that the

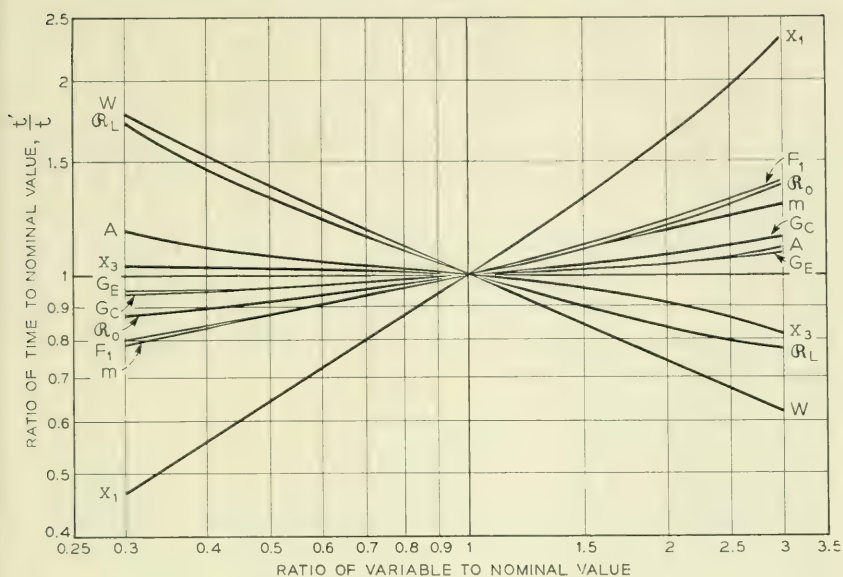


Fig. 9 — Mass controlled operate time variations.

relay always must operate. However, the purpose of these variation curves is to adjust measured data. The coil constant is the independent variable in any measurement and can always be set exactly. Hence it does not require an adjustment.

The chart shows clearly that the single most important factor is the armature travel. Following are power W , and leakage reluctance $\mathcal{R}_L$. The least sensitive is the pole face area A , because it has been optimized in the design. Taking the slopes at the (1, 1) point, the operate time, expressed in a separated variable form, becomes:

$$t_0 \approx x_1^{0.66} F_1^{0.24} \mathcal{R}_0^{0.2} m^{0.21} G_C^{0.84} G_E^{0.56} A^0 / W^{0.45} \mathcal{R}_L^{0.33} x_3^{0.84} \quad (28)$$

The exponents are a measure of sensitivity — the nearer they approach zero, the less the sensitivity. In designing electromagnets for speed, every effort should be made to keep the armature travel to a minimum as it is outstanding in its effect on time.

For our present purpose, the chart is used to adjust measured operate time data to the average value. The factors to be corrected are F_1 , m , x_1 , x_3 , A , $\mathcal{R}_0$, $\mathcal{R}_L$, i.e., all the pertinent geometric and load factors shown in Fig. 7. This completes the description of the method used for establishing the average mass controlled operate time curves.

Maximum and Minimum Values

The next to last part of this phase of the article concerns estimates of the range of the mass controlled maximum operate time of a particular relay code. This involves determining the range of each of the variables from the manufacturing specifications. Then with the variation chart the range in the time, due to each component, is evaluated, as a ratio. The actual range of course is not the product of these, as all variables never will be adverse simultaneously, but it is greater than the largest single individual contribution. For general purposes, a root sum square addition is used. For the wire spring relay this amounts to a range of about ± 30 per cent.

Preliminary Estimate of Type of Operation

Earlier, a rule of thumb expression for an estimate of whether an electromagnet is in the mass or load controlled class was given. It is easily derived by use of the $\dot{F}$ concept, and the observation from computing the variation charts, that the waiting time generally is about one-half of the motion time when the electromagnet is mass controlled.

When the motion time begins, the force starts to build up according to equation (19). However, multiplying this initial rate by the motion time need not result in a force value equal to the load for operation to be complete. Four factors act to this end: (1) during the waiting time the back tension F_1 already has been overcome; (2) when the armature gap closes, the armature takes a greater portion of the winding flux; (3) as the armature moves in the winding flux builds up more rapidly; and (4) the kinetic energy can cause the armature to crash through a small distance where the load projects through the rising dynamic pull curve.

Calling t'_M the longest average motion time which is mass controlled, and arbitrarily setting

$$\dot{F}_1 t'_M = \frac{F_3}{2}$$

in view of the foregoing, an estimation of the limiting average $\dot{F}_1$ is determined. Substituting this in (20), the maximum average mass controlled motion time is:

$$t'_M = \sqrt{\frac{12m(x_3 - x_1)}{F_3}},$$

which, when multiplied by $\frac{3}{2}$ to allow for waiting time, and dropping

the prime to indicate total time, becomes

$$t_M = \sqrt{\frac{27m(x_3 - x_1)}{F_3}}. \quad (29)$$

This equation was given as equation (1) in the early part of this article.

For the wire spring relay, this expression has a value of 7.5 millisecc. Increasing this by 30 per cent for a maximum value, an estimate of 10 millisecc results, agreeing with the earlier lower estimate of maximum time for load controlled operation.

Selection of Winding for Mass Controlled Operation

One more group of factors needs consideration before a winding is selected. These are (1) the range in dc resistance of the windings, (2) the winding temperature as determined by the duty cycle, and (3) the range in the battery voltage. The number of turns of the winding is ordinarily not considered as a variable once it is chosen because of the automatic machine method of winding. An examination of Fig. 6 shows that if the turns are too few, a greater time penalty obtains than if there are too many. Also, decreasing the circuit power, increases the best coil constant. These two considerations indicate that the best coil constant should be chosen under worst circuit conditions. For any other condition the operate time will be reduced. Further, the range between worst circuit and average time will be a minimum.

The procedure for choosing a winding is to determine the dc resistance of a maximum resistance winding at the operating temperature set by minimum battery voltage and the maximum duty cycle. This sets the worst circuit power, and by use of Fig. 6, the best number of turns. In no case is a winding specified with fewer turns than will supply sufficient ampere turns to operate the worst relay with the maximum load. In some cases of low power, this sets the number of turns. For some cases of intermediate power, heating requires the maximum winding surface area, also resulting in excess turns. The average resistance with its variation, all at a standardized temperature at 68°F, completes the design.

Summary of Single Relay Local Circuit Operation

An analytical determination of the operate time of a single relay cannot be obtained in closed form because it requires the solution of two simultaneous, non-linear, non-homogenous, differential equations, without adding the complications of representing the magnetic saturation

and eddy current effects in some convenient form. Approximations for solutions have been developed which predict with good accuracy the order of operate time attainable for a design, but not of sufficient accuracy to exactly determine the best winding for a particular case. Thus once a magnetic structure has been established, the operate time data for a single relay necessarily have to be determined empirically by measurements using controlled samples.

Actually this procedure is quicker and easier, and besides it gives the correct answer, including all the non-linear effects, such as eddy currents, saturation and motion, no matter how complicated. Complete data for the range of all relays and windings using the given structure can be determined using one such relay with a full winding, through the use of impedance transformation techniques and estimation of small variation effects, using the approximate solutions. By using the approximate solutions only for corrections, the errors become of second or smaller order. Thus single relay operate time data can be determined accurately and presented in a form permitting either analysis or synthesis of performance.

The design of a best winding for load controlled operation, is accomplished by a variation of the method of successive approximations. Appropriate battery voltage, winding temperature and resistance of course are considered as part of the solution of the problem. A range of windings is chosen and operate time data established for each winding. If necessary, the range is extended in the appropriate direction until a minimum operate time is included. This minimum determines the best winding.

For mass controlled operation, the contact spring load is immaterial, and the data can be presented directly. As part of this type of study, the relative importance of the several parameters affecting the operate time has been determined. These show at a glance whether (1) a change will have a significant effect or (2) what change or changes are necessary to effect a necessary reduction in operate time.

OPERATE TIME — SERIES RELAYS

Series relays are two or more relays whose windings all are connected in series, and energized by the same current, controlled by a single contact. The impedance of each one enters into the manner in which the common current will increase, after contact closure. If the procedure used for single relays were followed, there would be a double infinity of combinations to portray, or else experimentally study each combination

when proposed. This can be avoided, with no loss in generality, by transforming each relay into an equivalent single relay in a local circuit. Then the foregoing methods for single relays can be applied to each in turn.

This procedure is the common device of breaking up a complicated problem into parts, each of which can be solved by familiar methods. Two assumptions are made. The first is that each winding is a lumped two terminal network. At any instant the same current is in every turn of each relay and there is no propagation time involved. This holds for the times involved in electromagnets. The second assumption is that, when the relays have different operate times, the current reduction caused by the motional impedance of the first one to operate does not significantly extend the operate times of the later ones. In the following transformations, the winding turns, currents, and hence magnetomotive forces, are kept constant.

Identical Relays

If the two relays are identical they have some impedance $Z(p)$ which is the same for both. Part of this is the dc resistance and the other part is the ac effect, proportional to the turns squared. By the extension of Ohm's Law to ac circuits, when a potential source is applied to the two identical devices in series, exactly one-half the source appears across either device at any time after application, forming a virtual constant potential point. Thus if a battery of E_0 volts is applied, exactly one-half the battery appears across each, including the effect of eddy currents. Now if the voltage across a coil is known, then the response is uniquely determined, knowing just the relay characteristics and the voltage, disregarding the mechanism of how the voltage is applied. For this situation the voltage is in a most convenient form, represented by exactly one-half the battery. The operate time can easily be determined for either relay with this information, as the effects of eddy currents and motion are included in the data. The coil constant is already known and the power is one-half the total power.

General Case

This procedure can be generalized to include any division of dc resistance, different turns, and different magnetic structures providing, for the latter case, that eddy currents can be ignored. The justification for this will be considered later.

The basic problem is: given the battery voltage E_0 , relay No. 1 of N_1 turns, a 1 turn inductance L_1 , and resistance R_1 , in series with relay No.

TABLE I—EQUATION 30

Relay	Turns N	Resistance R_E	Voltage E_E	Power W_E	Coil Constant
No. 1	N_1	$\frac{R_1 + R_2}{1 + \frac{L_2}{L_1} \left(\frac{N_2}{N_1}\right)^2}$	$\frac{E_0}{1 + \frac{L_2}{L_1} \left(\frac{N_2}{N_1}\right)^2}$	$\frac{E_0^2}{(R_1 + R_2) \left[1 + \frac{L_2}{L_1} \left(\frac{N_2}{N_1}\right)^2\right]}$	$\frac{N_1^2}{R_1 + R_2} \left[1 + \frac{L_2}{L_1} \left(\frac{N_2}{N_1}\right)^2\right]$
No. 2	N_2	$\frac{R_1 + R_2}{1 + \frac{L_1}{L_2} \left(\frac{N_1}{N_2}\right)^2}$	$\frac{E_0}{1 + \frac{L_1}{L_2} \left(\frac{N_1}{N_2}\right)^2}$	$\frac{E_0^2}{(R_1 + R_2) \left[1 + \frac{L_1}{L_2} \left(\frac{N_1}{N_2}\right)^2\right]}$	$\frac{N_2^2}{R_1 + R_2} \left[1 + \frac{L_1}{L_2} \left(\frac{N_1}{N_2}\right)^2\right]$

2 of N_2 turns, a 1 turn inductance L_2 and a resistance R_2 . What are the two equivalent relays, each in a local circuit and what are their virtual applied voltages? The first step is to reassign the total resistance to obtain two equivalent windings having the same time constant $N^2 L_1 / R$. Then determine the two virtual voltages by division of the total in proportion to the impedances. These and their dependent power and coil constant relationships for the case of two relays are tabulated in Table I. The procedure can be extended to any number of dissimilar series structures.

For the case of all identical magnetic structures, the 1 turn inductances L_1 , L_2 , etc., are all equal and their ratios become unity, simplifying the expressions. Note that the coil constants are not equal unless the structures are the same magnetically, but that the time constants always are equal. Further, the total power is constant and equal to that of the original circuit.

After determining the effective powers and coil constants, the operate time for each can be read from the applicable single relay charts such as Fig. 6.

Two Like Parallel Relays in Series with a Third Relay, Identical Structures

The equivalent relay method can be extended to include the case of two like parallel relays in series with a third relay, as shown in Fig. 10. The first observation to make is that, because of symmetry, the current flow in the two parallel relays is identical. No winding current change would be made if, for instance, all the dc resistance of the two parallel relays were removed and half of either were connected in series with the single relay. Thus again, the resistances can be assigned as necessary to result in equal winding constants and total power. The same net dc resistance gives a second condition:

$$\frac{N_1^2}{R_{1E}} = \frac{N_2^2}{R_{2E}}, \quad (31)$$

$$R_1 + \frac{R_2}{2} = R_{1E} + \frac{R_{2E}}{2}.$$

Solving

$$R_{1E} = \frac{2R_1 + R_2}{2 + \left(\frac{N_2}{N_1}\right)^2}, \quad (32)$$

$$R_{2E} = 2 \left(R_1 + \frac{R_2}{2} - R_{1E} \right).$$

Because the equivalent coil constants have the same ratio as the equivalent resistances, the voltage division will be:

$$E_{1E} = \frac{R_{1E}E_0}{R_1 + \frac{R_2}{2}}, \quad (33)$$

$$E_{2E} = E_0 - E_{1E}.$$

With these, the equivalent coil constants and powers can be computed and charts such as Fig. 6 used as before.

Selection of Optimum Coils, Identical Structures

Neither Winding Known, Operate Times Given

In the above discussions it was assumed that the windings were known and that the operate time data were sought. The procedure can be reversed, starting with a desired operate time, or times, and then choosing optimum coils. From time curves such as Fig. 6, the watts corresponding to the desired times can be obtained provided they are read from the same coil constant vertical because the same current flows through both and hence the effective coil constants necessarily must be equal.

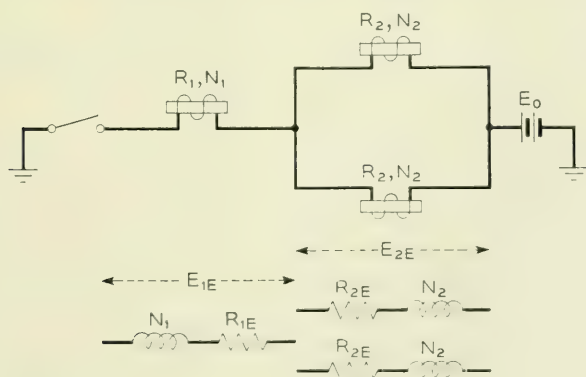


Fig. 10 — Transformation of series-parallel relay circuit.

The particular vertical N^2/R to be used is the one passing through the higher of the two given times, which lies at the minimum point on a watts curve. Satisfying the smaller wattage relay assures sufficient ampere-turns for both windings, and also gives a smaller time loss compared to optimum for the other higher speed relay, as the higher power curves are flatter for coil constants above the optimum. Thus the specified times determine equivalent watts for each coil, and the actual total circuit watts are exactly the sum of the equivalent watts. Hence:

$$W_{\text{total}} = \frac{E_0^2}{R_1 + R_2}. \quad (34)$$

This sets the total resistance ($R_1 + R_2$). Also, knowing the current is the same through the two coils, and that the actual division of the dc resistance has no effect on time, the ratio of the two resistances can initially be chosen the same as the ratio of the two effective powers:

$$\frac{W_1}{W_2} = \frac{R_1}{R_2}. \quad (35)$$

Finally, knowing that the effective coil constants for the two relays are equal, the desired turns are determined using the already selected coil constant, G_c . Solving these three equations, the individual relays are:

$$R_1 = \frac{W_1 E_0^2}{W_{\text{total}}^2}, \quad R_2 = \frac{W_2 E_0^2}{W_{\text{total}}^2}, \quad N_1 = \sqrt{R_1 G_c}, \quad N_2 = \sqrt{R_2 G_c}. \quad (36)$$

The resistance values R_1 and R_2 can be used as shown above or divided in any other way as long as the total is unchanged. The numbers of turns, however have to be kept fixed.

A particularly simple relationship exists when it is desired to have equal times. In this case the turns on the two coils are equal.

One Winding Known

As is often the case in actual use, a winding must be chosen to be used in series with another known winding and have best operate times. The method described here can be applied to any relay structure, but the numerical values in the analysis are applicable to the wire spring relays only. The first step is to choose the desired resistance for the coil. This is usually set by the heating and power limits, knowing that the higher the total power is, the less the operate time. Then the choice of turns for the coil depends on which of the two coils needs speed the most.

Assume first that we want speed on the coil to be designed, say relay No. 1, rather than the known relay in the circuit, say relay No. 2. Then we want the effective power for this relay as high as possible provided that the coil constant is not too far from optimum. From Table I it has already been noted that the two effective coil constants are always equal when the structures are identical. Also the sum of the two effective powers is exactly equal to the total power; that is, as the effective power to the first relay increases by increasing winding turns, that to the other correspondingly decreases. We see that for relay No. 1, the effective power increases as N_1 increases, but also that the effective coil constant increases. Thus an optimum turns value can be found, where on one side the low effective power slows up the operate time, and on the other the high coil constant does. Fig. 11 shows these optimum values. The solid curves for relay No. 1 are plots of time versus the turns ratio N_2/N_1 with total power, $E^2/(R_1 + R_2)$, and the "series coil constant" of relay No. 2, $N_2^2/(R_1 + R_2)$, as parameters. The optimum turns values for relay No. 1 show up clearly on this curve, and are seen to vary with both parameters.

This relation of optimum turns to the two parameters is shown on Fig. 12, where optimum N_2/N_1 values are plotted against total power with the relay No. 2 series coil constant as the parameter. The added series relay always has the most turns when it is designed for least time. Where speed on relay No. 1 is the only concern, the optimum turns can be chosen directly and easily from Fig. 12. Fig. 11, however is of more general use since it actually gives the times and also shows the time values for the second relay (the dotted curves). Thus the turns can be chosen to approach optimum speed on either relay or to choose a compromise value.

Now for relay No. 2 with total power $E^2/(R_1 + R_2)$, and the second relay series coil constant $N_2^2/(R_1 + R_2)$ as parameters, the effective power decreases and the effective G increases when N_1 is increased, both increasing the time. In other words, the given relay will always be slowed down by any added series relay winding.

The minimum N_1 value is limited by sufficient ampere-turns to operate the first relay. As shown by the dotted curves of Fig. 11, which are the time versus N_2/N_1 curves for relay No. 2, the gain in speed is slight as N_2/N_1 is increased beyond about 2. Thus, although a compromise must always be chosen for speed on relay No. 2, the loss in speed for relay No. 2 is not necessarily great if N_2/N_1 can be chosen near 2. For the case of equal operate times, in every case equality of turns of course applies.

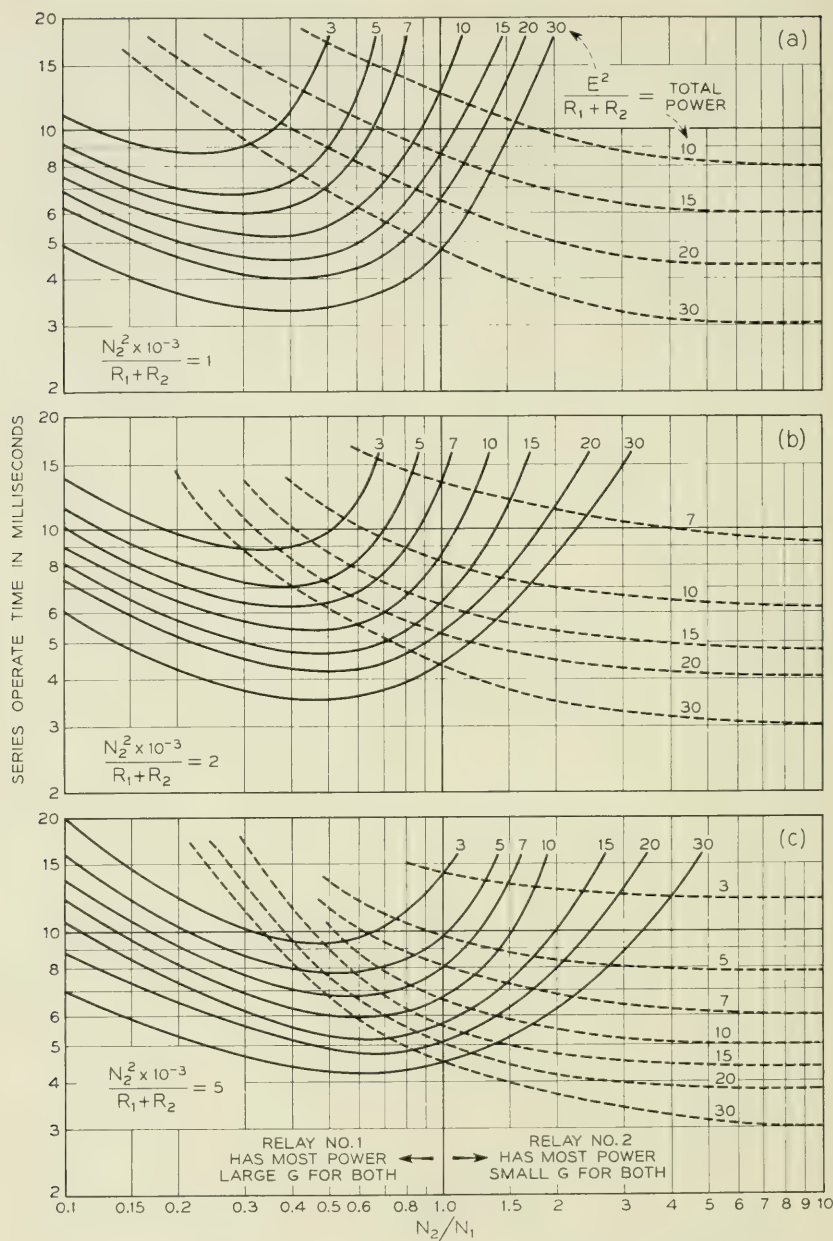
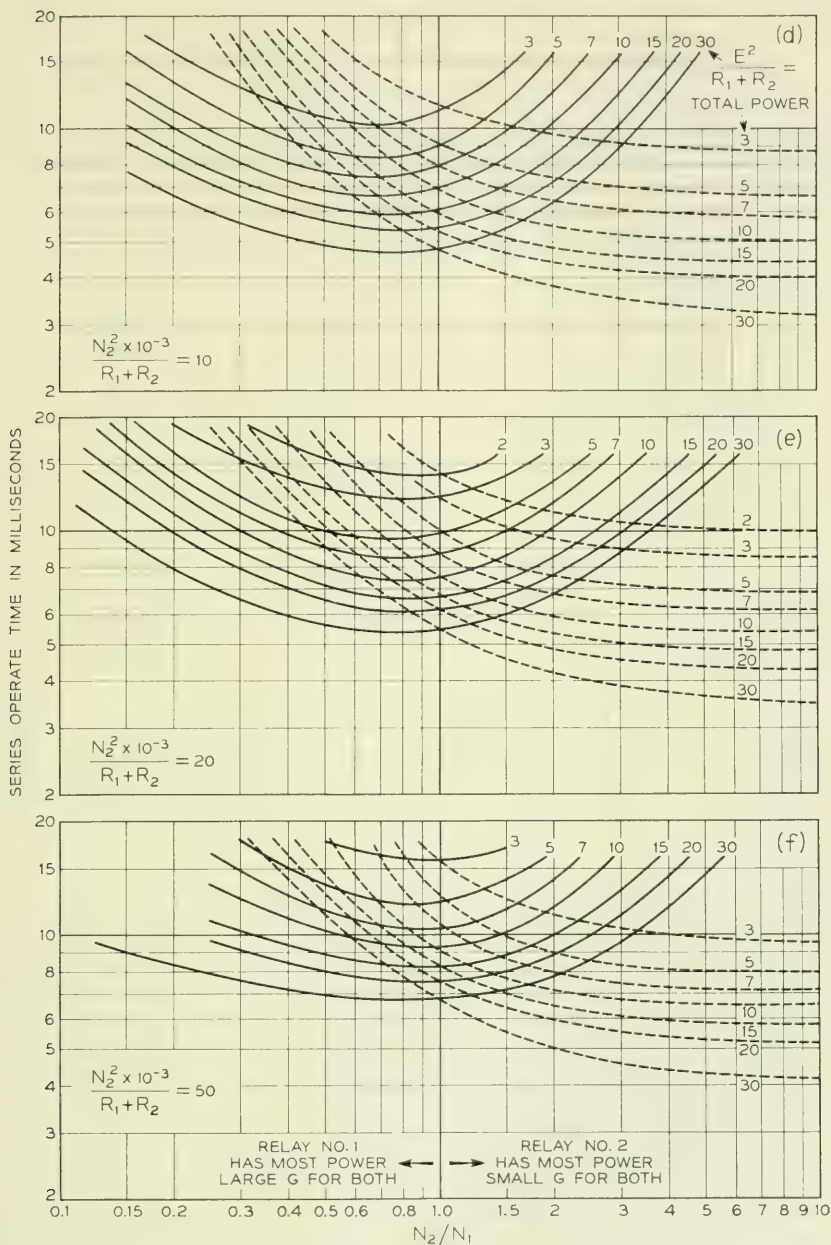


Fig. 11 — Operate time of



series relays versus turns ratio.

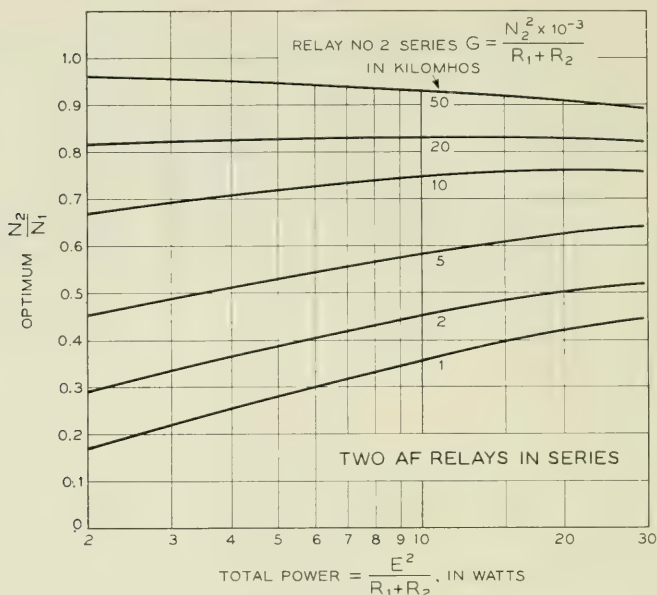


Fig. 12 — Optimum turns ratio for speed on relay No. 1, No. 2 known.

THE OMISSION OF EDDY CURRENT CORRECTIONS

The foregoing discussion of series connected relays developed transformation relationships ignoring eddy current effects. The following discussion will qualitatively show that ignoring the eddy current effects introduces second or smaller order errors.

For very large coil constants and different magnetic structures, the core constants can be ignored. With large coils, the current and flux rise are influenced very little by the core constant.² As the coil constant is reduced, the core effect becomes more important, but not significantly so until the speed class is reached. Unfortunately, we are generally only interested in the speed case. In the following, the most affected high speed applications are considered. It is found that the virtual constant potential point is only slightly affected by the eddy currents, and that equivalent single relays for each of the series relays determined using only the winding resistance and turns, with the relay reluctance correction as described above, are sufficiently accurate. The operate time for each relay itself, is affected by the core. This effect of course is included in the measured timing data which are used after equivalents are determined. The present discussion is directed toward the effect of the core eddy currents on the voltage division.

The linear equations used for developing the first approximation operate time equations also represent the behavior of the three element, two terminal network of Fig. 13, shown referred to a one turn admittance form. This exhibits the core eddy current effect as a resistor shunting an ideal inductance, rather than as an infinite line. Whatever value the shunting resistor may have, it affects only the transient response of the network, the part with which we are now concerned.

In what follows, it will be assumed that the transformation relations have already been applied, and the G_C values applying to Fig. 13 are effective values related to the winding turns in accordance with equations (30).

For two such networks in series, the voltage division would be independent of time if the ratios of the shunting to series conductances for each structure were equal. The method developed for similar structures then would apply with no error.

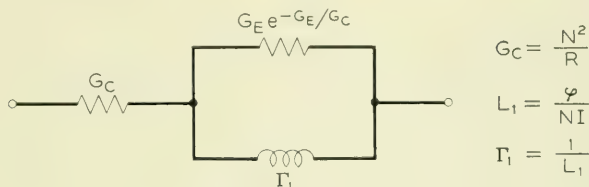


Fig. 13 — Equivalent linear circuit represented by time equation.

The ratio is:

$$\frac{G_E}{G_C} e^{-G_E/G_C} = a e^{-a}.$$

This function has a maximum value of 1 e ; that is, the shunting resistor is at least e times larger than the series resistor. An analysis of the range of core conductances and speed windings in use, shows that the actual resistance ratios are somewhat greater than this and hence the ratios are not quite independent of the structure. However, because the maximum is broad, it reduces the actual range to about 10 per cent, including all windings plus a 2 to 1 G_E change. In turn, this signifies that for a suddenly applied voltage, the initial linear network voltage division would differ from the final by less than 10 per cent. This error decreases with time.

The above discussion applies to linear networks such as Fig. 13. In an actual magnetic structure the initial voltage division is not affected by eddy currents as they have not had time to build up. The voltage

division starts and ends exactly in the ratio of the effective series resistances by virtue of the method used in determining them. For times after circuit closure, a transient voltage error does develop, but it will not approach in magnitude the initial transient error of the linear network. The operate time error will be still smaller. The operate time varies as the two-thirds power of the voltage (power varies as the voltage squared).

These considerations lead to the conclusion that ignoring the eddy currents in setting up the equivalent relays, results at the most in errors of operate time estimates for series connected relays of the order of a few per cent.

RELEASE TIME

The release time of a relay is not as directly affected by the winding, as is the operate time. For instance, if it is opened by the controlling contact without an RC network, the winding current, ignoring arcing at the contacts, abruptly drops to zero. The winding subsequently plays no part in the flux decay. If the winding has a shunting resistor or RC circuit, then winding current does flow and some effect is present. A resistor always increases the release time. A favorable RC choice can cause a slight decrease in release time. Because of this minor effect, the winding almost always is designed from operate time or sensitivity considerations.

Differing slightly from operate time, the release time is divided into two parts, (1) waiting time plus motion time until actuation of the nearest contact and (2) stagger time. For simplicity the combination (1) above is merely called release waiting time.

The release time of a relay is a more complicated function than is the operate time. The primary cause of this is the closed gap situation, with little stabilizing effect from an air gap. For the earlier operate time studies the air gap largely contributed to the simple exponential relationships. A second effect in release is that the magnetic material is almost always in the non-linear saturated portion of its characteristic. Hence an approximately linear relationship between steady state flux and current cannot be assumed. Finally, for release without an RC network, the only current flows in the core, so it completely controls the flux decay. Even with an RC network, only a minor decrease in release time is possible. For these reasons, except for rough preliminary estimates of release times during preliminary design, all release data are based on measurements.

Conductance Shunt

In a companion article,³ a hyperbolic relationship between the flux and current is used to represent the portion of the hysteresis loop of concern in release. This, with a conducting sleeve or shunted winding, gives an excellent representation of the release waiting time. It also provides an understanding of the controlling parameters, even with only eddy currents controlling the release. The form most useful for release consideration is:

$$t = \frac{(\varphi'' - \varphi_0)(G_s + G_c + G_E)}{NI_0} \left(\frac{\ln z}{z - 1} - \frac{1}{z} \right), \quad (37)$$

where

$$z = \frac{\varphi'' - \varphi_0}{\varphi - \varphi_0}.$$

In these expressions, φ is the flux corresponding to the ampere turns NI_0 at which the relay just releases, φ_0 is the residual flux, and φ'' is the asymptotic saturation flux. The conductance terms are the same as for the operate case except G_c now is computed using the dc winding resistance plus any winding shunting resistance. If the winding has no shunt then $G_c = 0$. The operated contact spring load enters through φ and NI_0 .

The function of z in brackets has a broad maximum in the region of relay release. It therefore is appropriate to consider the releasing ampere turns as the independent variable and plot the measured release time with the total conductance as a parameter covering the range of interest. As G_E is a characteristic of the structure and not subject to adjustment, the curves are actually labelled in terms of just the sleeve, if any, plus the shunted winding. For the wire spring relay, Fig. 14 shows data in this form. If the function of z were truly constant, the curves would all have a slope of -1 in this plot.

Strictly speaking, the above equation applies only to the time until the magnetic pull has decayed to equality with the operated spring load. Following this is the motion time, during which the pull decays further. For convenience, however, the armature motion time through the distance to the nearest contact is included in the chart for release time as (a) the contact actuation is the means used to measure the time and (b) this much motion always takes place before any contact is actuated.

The further displacement of the armature continues at almost con-

stant velocity. This is because the decreasing magnetic pull (called drag for the release case) has become small and the contact spring load is dropped by the motion. The further forces acting on the armature then are only the back tension and the difference between the rapidly decreasing contact spring load and magnetic drag.

Returning now to equation (37) above, the release time is directly proportional to the sum of the conductances, and to the difference between the saturation and residual fluxes. The conductance variations can be determined from temperature, winding data, sleeve dimensions, and material. The saturation flux is directly proportional to the core

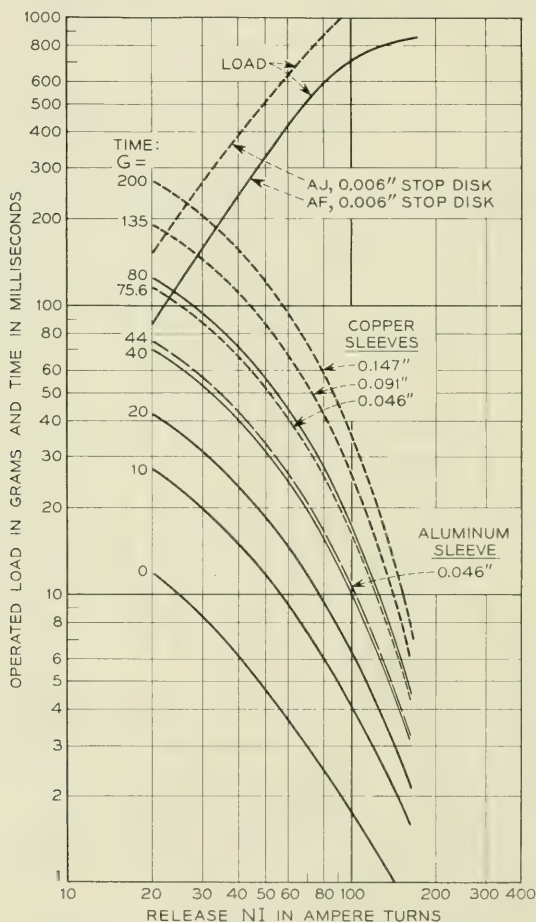


Fig. 14 — Release time with a shunted winding.

cross-section. The residual flux is directly proportional to the coercive force times the length of the magnetic material, and inversely proportional to the closed gap reluctance. The two largest factors affecting the closed gap reluctance are magnetic permeability, and fit of the joints, including variations in stop pin height. By measuring these factors on the test relay and estimating the corresponding values for the desired reference condition, the measured data can be corrected to the reference condition. Then having the release pull curves after magnetic soak for the same reference condition, the release ampere turns for any load under consideration and hence its release waiting time can be determined. These release pull curves are also shown as part of Fig. 14 for the wire spring relay. The ordinate scale is marked for both contact spring load in grams and releasing time in milliseconds. The particular chart shown is for a constant initial number of ampere turns. For other initial values, a correction chart is provided. If not as desired, the time can be adjusted either upward or downward by changing to a different sleeve, shunting resistor, or both. Of course, in no shunting conductance case can the release time be less than the open circuit time.

RC Shunt

The above considerations all related to shunting conductances. These serve the purpose of increasing the release time. The shunting resistor also greatly reduces the transient peak voltage developed when the winding circuit contact is opened. For contact protection reasons, RC networks are frequently used across the winding or contact for this same purpose. Such a network also has no power drain when the relay winding is energized. By a suitable choice of capacitance the network can reduce the release time to a value less than the open circuit time. It does this by developing a heavily damped oscillation of winding current. The frequency must be of the order of the reciprocal of the unprotected release time for a release time reduction. This fixes the choice of capacitance to a small range i.e., there is a best capacitor, one for each winding. Smaller values cause the time to increase toward the unshunted case. Larger values also cause an increase but for this case, if too large, an increased time beyond the unshunted case can result.

For speed windings, where timing is important, a network is designed for each. For the slower higher resistance windings, because time is not as important, the closest to the best of the available networks is chosen. This results in a time penalty but furthers standardization.¹

To evaluate a network, the transformations developed and shown in

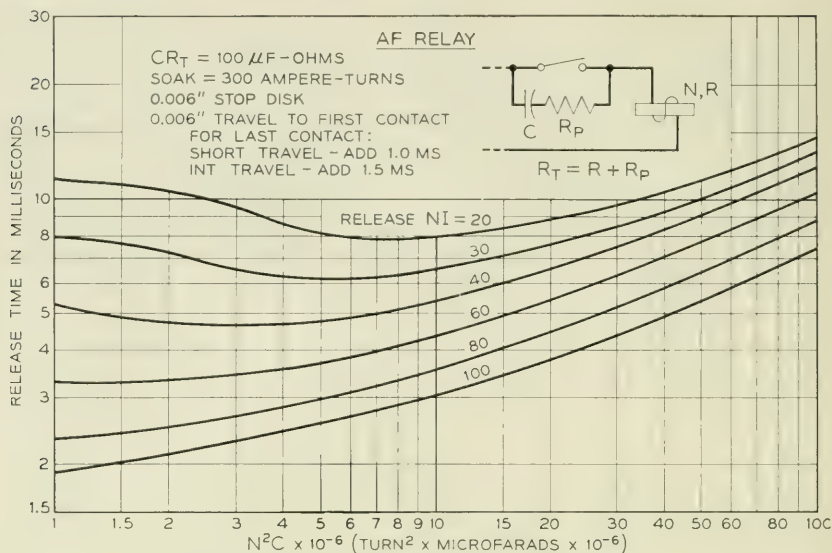


Fig. 15 — Release time as a function of N^2C .

Fig. 5 are used. This permits exact simulation for any winding and network from a time standpoint. For presentation of data, an arbitrary separated product form for the more important variables is used. The particular factors chosen are:

(1) N^2C , a measure of frequency, with NI_0 (RELEASE NI) as a parameter.

(2) RC , a damping term, with NI_0 as a parameter.

(3) Soak NI with N^2/R as a parameter.

Essentially what is assumed is that the release time can be represented well enough as a function of the form:

$$t_1 = f_1(N^2C) \times f_2(RC) \times f_3(NI). \quad (38)$$

Then by holding two factors fixed and varying the third, each factor in turn is evaluated. Generally the accuracy is better than 10 per cent even with this elementary approach. For a few cases the error approaches 15 per cent. Typical charts for these three factors are shown in Figs. 15, 16 and 17.

Release Time for Series Relays and an RC Shunt

With two series relay windings it is usual to provide a single RC network, as shown in Fig. 18. This results in a double infinity of possible combinations if handled directly. However, it can be broken into two

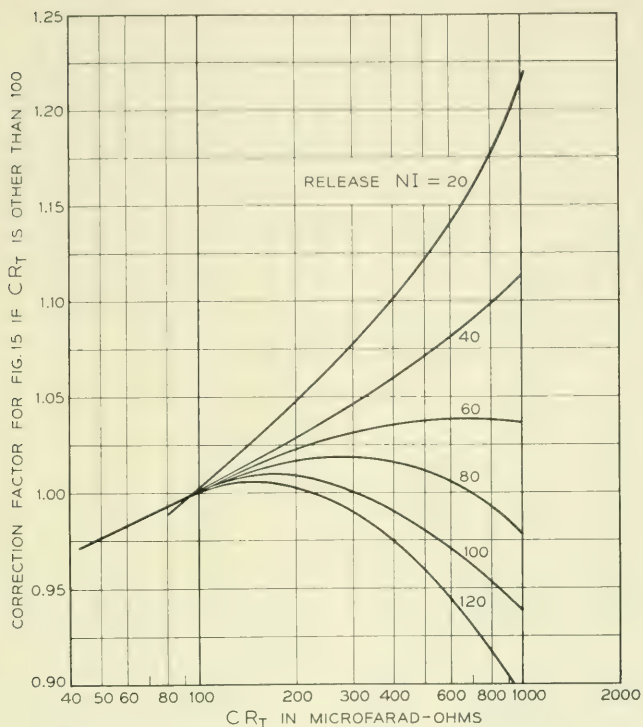
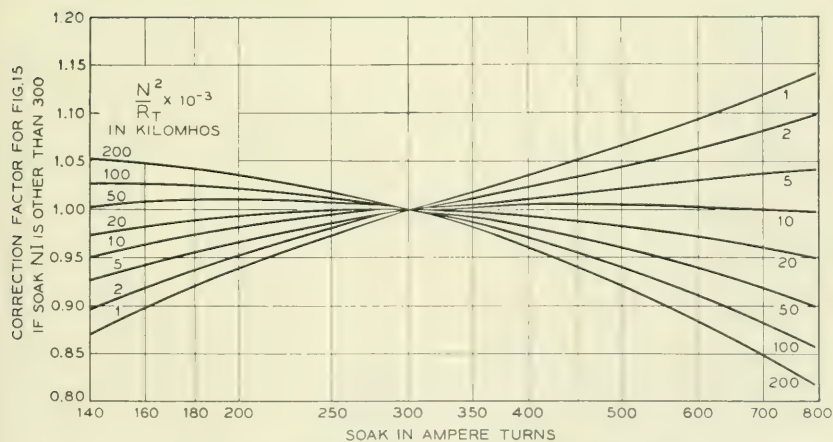
Fig. 16 — Release time correction for RC damping.

Fig. 17 — Release time correction for magnetic soak.

parts, each of which exactly represents, as individual relays, the two series relays. This is done by first assigning the total of the winding resistances for equal N^2/R values. The two equivalent relays then form a voltage divider, independent of frequency. The RC network next is drawn as two series RC networks as shown in Fig. 19. The procedure now is to determine the component RC values to have the identical voltage divider effect as the relays. Then the two equal voltage points can be connected as shown by the dotted line and no circulating current will flow. The resultant is a three node circuit and each branch can be considered independently, using the method of the preceding section.

There are four unknowns and hence four equations are needed. They are

$$\begin{aligned} R_{11} + R_{22} &= R, & \frac{N_1^2}{R_{11}} &= \frac{N_2^2}{R_{22}}, \\ \frac{C_1 C_2}{C_1 + C_2} &= C, & N_1^2 C_1 &= N_2^2 C_2. \end{aligned} \quad (39)$$

The solutions for the four unknowns are shown on Fig. 19.

For experimental measurements, each is transformed again by means of Fig. 5. Note that the time constants all are equal, as they must be, because the same current flows through all elements. However, the initial ampere turns NI are different by virtue of $N_1 \neq N_2$. Thus only one circuit like Fig. 5 need be set up. Measurements then are made with two different voltages to provide the two different ampere turns and the two release times. By choosing the subscript 1 network to represent the actual circuit, then no subscript confusion results in arriving at the simulating circuit of Fig. 5.

Release Time for Similar Parallel Relays and an RC Shunt

Similar parallel relays are split as shown in Fig. 20. Two equal parallel RC networks are first drawn. One is then assigned to each, and then the pairs are divided. The release times are each equal to that of one of the separated circuits.

Summary

The release time of fast electromagnets is influenced much more than the operate time by the fit of the magnetic parts. For release, the small non-magnetic stop disc introduces a relatively small stabilizing air gap compared to the open gap of the operate case. Secondly, the release always starts after an applied magnetomotive force which differs as

between circuits, whereas operation invariably follows an open circuit. Thirdly, with RC protection networks connected, a more complicated winding current is present. Finally, the transmission line behavior of the magnetic core material is less masked by the winding, and in fact controls completely for the open circuit case. For these reasons, the

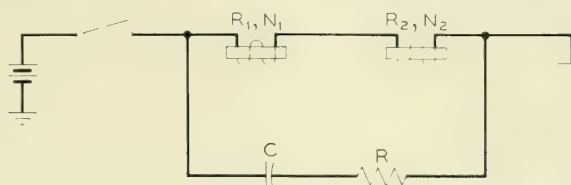
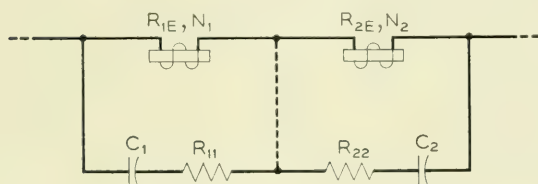


Fig. 18 — Series relays with one RC network.



$$R_{11} = \frac{R}{1 + \left(\frac{N_1}{N_2}\right)^2} \quad C_1 = C \left[1 + \left(\frac{N_2}{N_1}\right)^2 \right]$$

$$R_{22} = \frac{R}{1 + \left(\frac{N_2}{N_1}\right)^2} \quad C_2 = C \left[1 + \left(\frac{N_1}{N_2}\right)^2 \right]$$

Fig. 19 — Transformation to series RC networks.

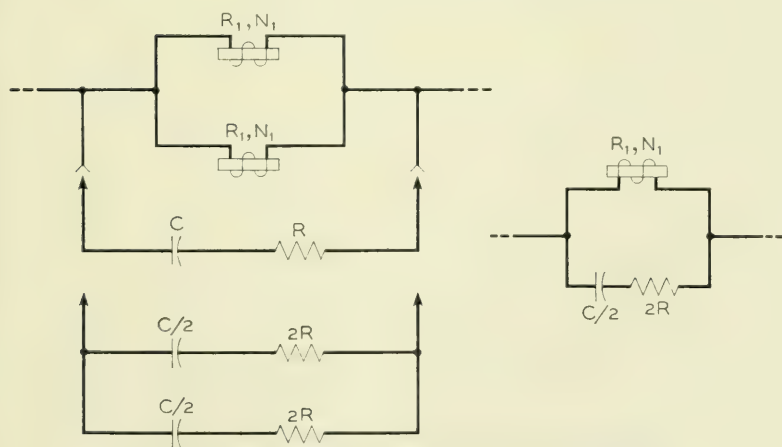


Fig. 20 — Transformation to parallel RC networks.

analytical presentation, as noted in Part I, of fast release time data is not now as advanced as the operate time data.

The material presented here describes the present state of the art. For slow releasing relays, the performance can be predicted with accuracy. For fast relays, while the general pattern is known, accurate means for estimating variations have not been developed. The present engineering of releasing relay circuits therefore, depends upon specific measured data in chart form, for each condition.

ACKNOWLEDGMENTS

The analyses leading to the forms of data presentation for the several types of relay timing information are the results of contributions from many people. In particular, the early nomograms for load controlled operation were developed by P. W. Swenson. The node method for separation of series releasing relays shunted by an RC network was developed by R. H. Gumley. The graphical method for design of optimum series windings was developed by Mrs. K. R. Randall. To her, I am also indebted for preparation of all the charts of measurements.

REFERENCES

1. H. N. Wagar, Economics of Telephone Relay Applications, page 218 of this issue.
2. M. A. Logan, Dynamic Measurements on Electromagnetic Devices, B.S.T.J., **31**, pp. 1413-1466, Nov., 1953.
3. R. L. Peek, Principles of Slow Release Relay Design, page 187 of this issue.
4. R. L. Peek and H. N. Wagar, Magnetic Design of Relays, page 23 of this issue.

Principles of Slow Release Relay Design

By R. L. PEEK, Jr.

(Manuscript received September 25, 1953)

This article presents an analytical treatment of the relations controlling the release time of slow release relays, in which field decay is delayed by the currents induced in a conducting sleeve or slug. An hyperbolic relation between flux and magnetomotive force, fitting the decreasing magnetization curve, is used for the relation between induced voltage and current in the sleeve circuit in determining the rate of field decay. Methods are given for estimating and measuring the magnetization constants and those appearing in the relation between pull and field flux.

These relations are used as a basis for a discussion of the design of slow release relays and of the adjustment procedures employed to meet timing requirements.

1 INTRODUCTION

Slow release relays are built and adjusted to provide a time delay between the opening of the coil circuit and the release motion which restores the contacts to their unoperated condition. They are used to assure a desired sequence of circuit operation, as, for example, in maintaining a closed path through the slow release relay's contacts during the pulses sent in dialing a digit, and opening this path during the much longer interval between digits. Slow release relays constitute a minor but significant part of the relay population in an automatic central office: about ten per cent of the total.

For economy in manufacture, installation, and use, slow release relays are made as similar to the ordinary or general purpose relays with which they are used as their special requirements permit. Each general structural relay developed for telephone work has had a variant form for slow release use. Thus the *Y* type¹ relay is the slow release form of the *U* type relay² widely used in the Bell System, while the *AG* relay is the slow release variant of the recently developed wire spring³ relay.

The circuit functions of most slow release relays permit considerable variation in release time if the minimum delay specified is assured. This

tolerance in the timing requirements permits the use of more economical practices in the construction and use of slow release relays than would be needed for closer control. As these wide tolerances apply to the great majority of applications, over-all economy is attained by developing slow release relays with reference to them, using other devices for the special applications requiring close timing control.

The magnitude of time delay desired is fixed by the operate and release times of the associated general purpose relays. The latter times lie in the range from 5 to 50 milliseconds, with the majority in the lower part of the range. The delays required to assure sequences of events each requiring time intervals of this order consequently cover the range from 50 to 500 milliseconds (one twentieth to one half a second). Delays of less than 100 milliseconds can generally be provided by ordinary general purpose relays having shorted secondary windings or sleeves (single turn conductors) to delay their release. Special slow release relays are used for delays in excess of 100 milliseconds.

Factors Controlling Release Delay

In terms of circuit operation, release time is the interval from the opening of the coil circuit to the completion of contact actuation during the return motion of the armature. This time is the sum of (a) the time for the magnetic field to decay to the level at which the pull just equals the operated spring load and (b) the motion time for contact actuation. The motion time is never more than a few milliseconds, and is therefore a trivial increment to the long delays of slow release relays. The release time of such relays is therefore, for practical purposes, simply the time of field decay.

When the coil circuit is opened, the field decay induces currents in any circuit linking the field. These currents tend to maintain the field and to delay its decay. A short circuited winding or the single high conductivity turn provided by a copper sleeve gives a high magnetomotive force for a low induced voltage, resulting in a relatively slow decline in field strength. The time for the flux to reach a given level depends upon the conductance of the sleeve, and upon the reluctance of the electromagnet: the ratio of magnetomotive force to flux. The flux level which determines the end of the delay is that for which the pull equals the spring load. The delay is therefore prolonged by a pull characteristic such that the load is held until the flux drops to a minor fraction of its initial value.

Thus the essential features of a slow release relay are a shorted winding

or sleeve, a low reluctance magnetic circuit, and a high level of pull for relatively low values of field strength. In the following analysis of slow release performance, expressions are developed for the time of field decay for the decreasing magnetization characteristic, for the reluctance of the electromagnet, and for its pull characteristics. These expressions permit the estimation of release times, indicate the design conditions to be satisfied to attain a desired level of delay, and indicate the effect on the release time of variations in the spring load, in the dimensions of the electromagnet, and in other design parameters.

The notation used in this article conforms to the list that is given on page 257.

2 FIELD DECAY RELATIONS

If a closed circuit of resistance R and N turns links a magnetic field of flux φ , the voltage equation is:

$$iR + N \frac{d\varphi}{dt} = 0,$$

where i is the circuit current. Multiplying by $4\pi N$ and dividing by R this equation may be written as:

$$\mathfrak{F}_i + 4\pi G_i \frac{d\varphi}{dt} = 0,$$

where $\mathfrak{F}_i$ is the magnetomotive force of this circuit, and G_i is its value of N^2/R , which may be termed the equivalent single turn conductance. If there are several such circuits linking the same magnetic field, a similar expression applies to each, and these may be added to give the equation:

$$\mathfrak{F} + 4\pi G \frac{d\varphi}{dt} = 0, \quad (1)$$

where $\mathfrak{F} = \sum \mathfrak{F}_i$, and $G = \sum G_i$. In the case of a slow release relay, one linking circuit is usually a sleeve, whose conductance may be designated G_s . The applications are identical for a short circuited winding, if the applicable value of N^2/R is substituted for G_s . In either case, G also includes a term G_E representing the net effect of the eddy current paths. As the eddy currents at different distances from the center of the core link different fractions of the total field, this representation by a single term is an approximation. As shown in a companion article,⁴ however, the approximation is satisfactory when G_E is a minor part of G , as in the

slow release case. In this same article it is shown that, under these conditions, the relation between the changing values of $\mathfrak{F}$ and φ in (1) is substantially identical with that given by static magnetization measurements.

The static magnetization relation between $\mathfrak{F}$ and φ may therefore be substituted in (1), which may then be integrated to give the relation between φ and t . For linear magnetization, or constant reluctance $\mathcal{R} = \mathfrak{F}/\varphi$, integration of (1) gives the equation.

$$\frac{\varphi}{\varphi_1} = e^{-t/t_G},$$

where φ_1 is the value of φ for $t = 0$, and $t_G = 4\pi G/\mathcal{R}$, the time constant for this simple exponential decay. For release, however, the magnetization relation is not linear, but has the character shown in Fig. 1. The φ versus $\mathfrak{F}$ relation is asymptotic to the saturation flux φ'' , and has an intercept φ_0 at $\mathfrak{F} = 0$, resulting from the remanence of the magnetic material. The relation between φ and t in release is given by (1) with the relation between φ and $\mathfrak{F}$ that of the decreasing magnetization curve illustrated in Fig. 1.

Graphical Determination of Release Time

Writing $4\pi Ni$ for $\mathfrak{F}$ in (1), the integral form of this equation is:

$$\frac{t}{G} = \int_{\varphi}^{\varphi_1} \frac{d\varphi}{Ni}, \quad (2)$$

where t is the time for the field to decay from an initial value φ_1 to a final value φ . If experimental data for the decreasing magnetization curve are available, the integral of (2) may be evaluated graphically as is indicated in Fig. 2.

For purposes of illustration, the left hand plot shows two decreasing magnetization curves, such as might be obtained with two models of the same relay. The dashed curve 1 corresponds to a higher value of φ_0 and a lower reluctance than apply to the solid curve 2. To evaluate the integral of equation (2), the curves are replotted as at the right to give φ versus $1/(Ni)$. Then the integral of equation (2) is given by the area between the curve and the axis of φ . For curve 1, for example, the area a - d - a' measures the value of t/G for the flux to decay from φ_1 to the value of φ at a .

If the areas are evaluated for a series of points, such as a and c , the resulting values of t/G can be plotted, either against the correspond-

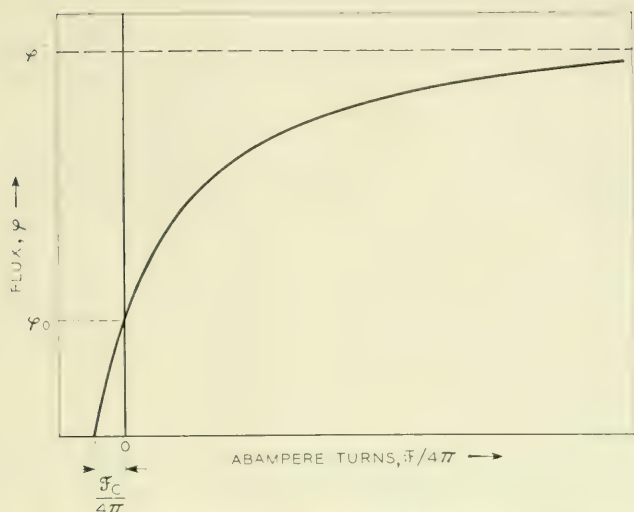


Fig. 1 — Decreasing magnetization of an electromagnet.

ing values of φ , as in Fig. 3, or against the corresponding values of NI , as in Fig. 4. In comparing two models, corresponding to curves 1 and 2 in these figures, the curves of $t'G$ versus φ in Fig. 3 indicate the comparative release times for the same spring load, as the pull versus φ relation is the same for similar structures. Thus in Fig. 3 the times for the pull to drop to a common value for curves 1 and 2 correspond to the points c and b , for example, and are measured by the areas $o-c-c'$ and $o-c-b'$ respectively. Hence curve 1 gives longer times in this comparison than curve 2, as illustrated in Fig. 3.

If, on the other hand, the spring loads for the two models are adjusted so that they release on the same value of NI , the release times to be compared correspond to decay to the points a and b , for example. The corresponding values of $t'G$ are given by the areas $o-d-a'$ and $o-c-b'$. These differ little from each other, and consequently curves 1 and 2 in Fig. 4 are similar. It is therefore possible to adjust individual relays to give nearly the same release time by adjusting the spring load so that they release on the same value of NI . Advantage is taken of this fact in the adjustment practices employed with slow release relays, as discussed in a latter section.

This graphical method for the prediction of release times, which was developed by H. N. Wagar,⁵ is described here because of its utility in demonstrating the relation between the timing relations and the character of the demagnetization curves. It provides a means for accurately predicting the release times in specific cases, but does not afford as

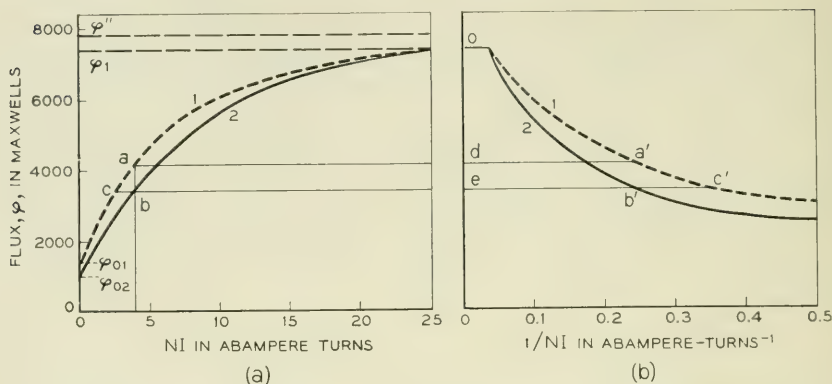


Fig. 2 — Graphical determination of release time.

explicit a statement of the general relations applying as does the approximate analytical treatment outlined below.

Hyperbolic Approximation to Decreasing Magnetization Curve

The decreasing magnetization curve, as illustrated in Fig. 1, has the general character of a rectangular hyperbola, and may therefore be represented approximately by the empirical equation:

$$\frac{\mathcal{F} + \mathcal{F}_c}{\phi} = \frac{R''\phi''}{\phi'' - \phi}, \quad (3)$$

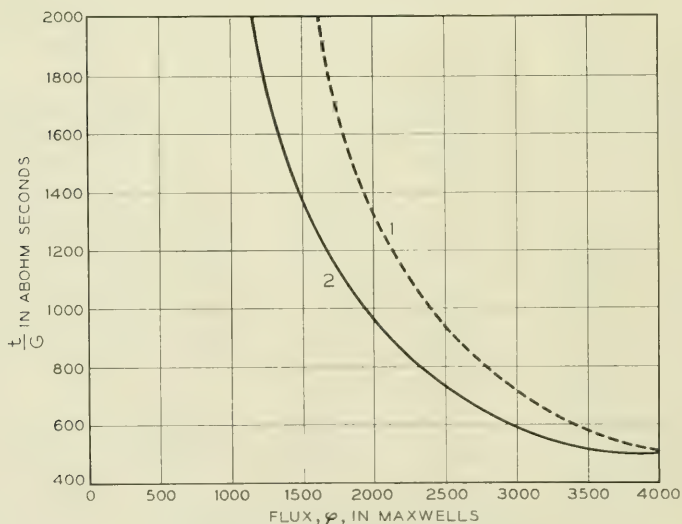


Fig. 3 — Release time versus flux.

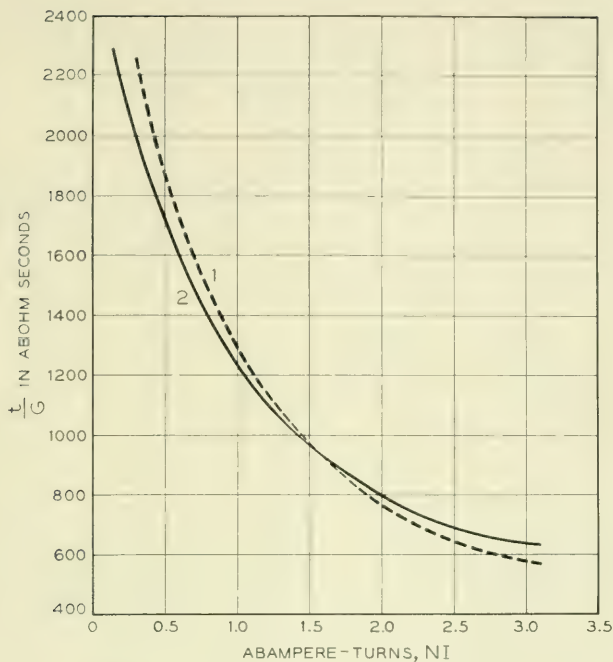


Fig. 4 — Release time versus ampere turns.

for which $\varphi = 0$ for $\mathfrak{F} + \mathfrak{F}_c = 0$, while φ'' is the asymptote approached by φ as $\mathfrak{F}$ is increased. $\mathcal{R}''$ is the initial incremental reluctance, the value of $d\mathfrak{F}/d\varphi$ for $\varphi = 0$. If φ_0 , as in Fig. 1, is the value of φ for $\mathfrak{F} = 0$, the resulting expression for $\mathfrak{F}_c$ may be substituted in (3) and this equation written in the alternative form:

$$\frac{\mathfrak{F}}{\mathcal{R}''\varphi''} = \frac{\varphi}{\varphi'' - \varphi} - \frac{\varphi_0}{\varphi'' - \varphi_0}. \quad (4)$$

The incremental reluctance $\mathcal{R}_i$, the value of $d\mathfrak{F}/d\varphi$ at $\varphi = \varphi_0$, is given by:

$$\mathcal{R}_i = \left(\frac{\varphi''}{\varphi'' - \varphi_0} \right)^2 \mathcal{R}'' . \quad (5)$$

The expression for $\mathfrak{F}$ (4π) given by (4) may be substituted for Ni in (2) to give an expression for the release time t . After clearing fractions, the resulting equation is:

$$\frac{t}{G} = \frac{4\pi(\varphi'' - \varphi_0)^2}{(\mathcal{R}''\varphi''^2)} \int_{\varphi}^{\varphi_1} \left(\frac{1}{\varphi'' - \varphi_0} - \frac{1}{\varphi - \varphi_0} \right) d\varphi.$$

From (5), the term outside the integral sign is $4\pi'(\mathcal{R}_i)$. On integration, the following expression is obtained for the release time t :

$$t = \frac{4\pi G}{\mathcal{R}_i} \left(\frac{\varphi - \varphi_1}{\varphi'' - \varphi_0} + \ln \frac{\varphi_1 - \varphi_0}{\varphi - \varphi_0} \right).$$

In slow release relays, a "soak" or high ampere turn value is applied in operation, and the initial value φ_1 is close to the saturation value φ'' . It is therefore a satisfactory approximation to take $\varphi_1 = \varphi''$ in the preceding equation, which then reduces to:

$$t = \frac{4\pi G}{\mathcal{R}_i} \left(\ln z - 1 + \frac{1}{z} \right), \quad (6)$$

where,

$$z = \frac{\varphi'' - \varphi_0}{\varphi - \varphi_0}. \quad (7)$$

The time therefore varies as the bracketed term in (6), which is shown plotted against z in Fig. 5. As z varies inversely as $\varphi - \varphi_0$, the difference between the flux and its ultimate value, (6) gives the release time for the value of φ at which the pull equals the operated load.

To obtain an expression for the release time in terms of the ampere turn value at which release occurs requires an expression for z in terms of $\mathfrak{F}$, or $4\pi NI$. This may be obtained by substituting in (4) the expression for $\mathcal{R}''$ given by (5). The resulting equation reduces to:

$$\mathfrak{F} = (\varphi'' - \varphi_0)\mathcal{R}_i \frac{\varphi - \varphi_0}{\varphi'' - \varphi} = \frac{(\varphi'' - \varphi_0)\mathcal{R}_i}{z - 1},$$

giving the equation:

$$z = 1 + \frac{(\varphi'' - \varphi_0)\mathcal{R}_i}{\mathfrak{F}}. \quad (8)$$

If $\mathfrak{F}$ is the value of $4\pi NI$ at which release occurs, the corresponding value of z given by (8) may be substituted in (6) to determine the release time. In this way there have been determined the values of $t\mathcal{R}_i/(4\pi G)$ plotted against $\mathfrak{F}/(\mathcal{R}_i(\varphi'' - \varphi_0))$ in Fig. 6. This is a universal curve for the relation between release time and the ampere turn value at which release occurs. The observed relation for any specific case is given by this curve, displaced vertically by the value of $4\pi G'(\mathcal{R}_i)$ and horizontally by the value of $\mathcal{R}_i(\varphi'' - \varphi_0)$ for the case in question. This is illustrated in Fig. 7, which shows the observed release time versus release ampere turn curves for the Y type relay for two sleeve sizes, corresponding to

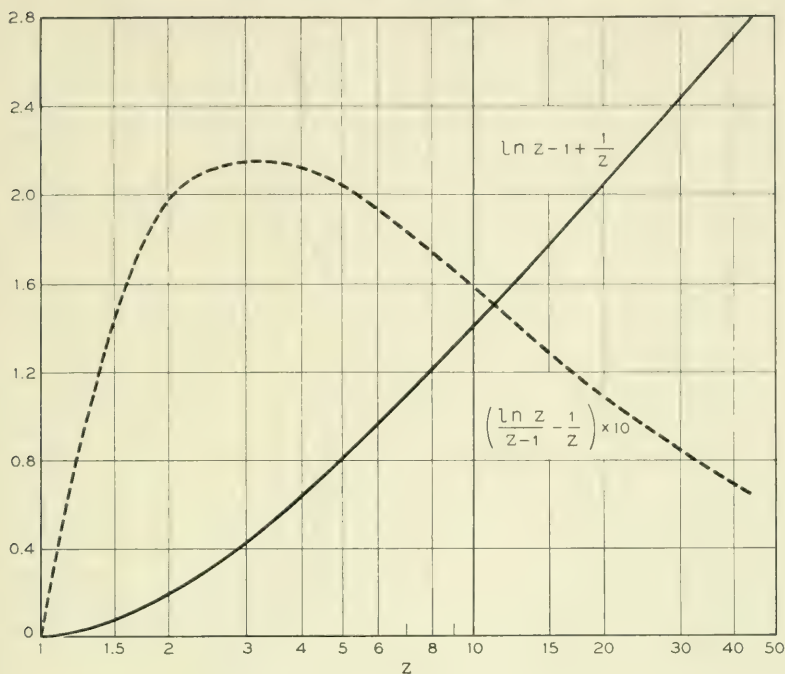


Fig. 5 — Factors for computing release time.

different values of G . For each sleeve size two curves are shown, corresponding to the limits of variation in magnetic characteristics, or in $\mathcal{R}_i$ and $\varphi'' - \varphi_0$. The dotted curve included for comparison is the relation of Fig. 6.

An important property of the t versus $\mathfrak{F}$ relation can be demonstrated by substituting in (6) the expression for $\mathcal{R}_i$ given by (8). There is thus obtained the equation:

$$t = \frac{4\pi G(\varphi'' - \varphi_0)}{\mathfrak{F}} \left(\frac{\ln z}{z-1} - \frac{1}{z} \right). \quad (9)$$

A plot of the function of z appearing in brackets in this equation is included in Fig. 5. In the vicinity of the maximum at $z = 3.09$, this function is nearly a constant, varying little for values of z between 2 and 6. In this range t varies inversely with $\mathfrak{F}$, and the proportionality constant depends only upon G , the sleeve conductance, and the term $\varphi'' - \varphi_0$, which is relatively independent of material and dimensional variations, as shown later. This confirms the conclusion, previously indicated by graphical analysis, that through a considerable range of operation the

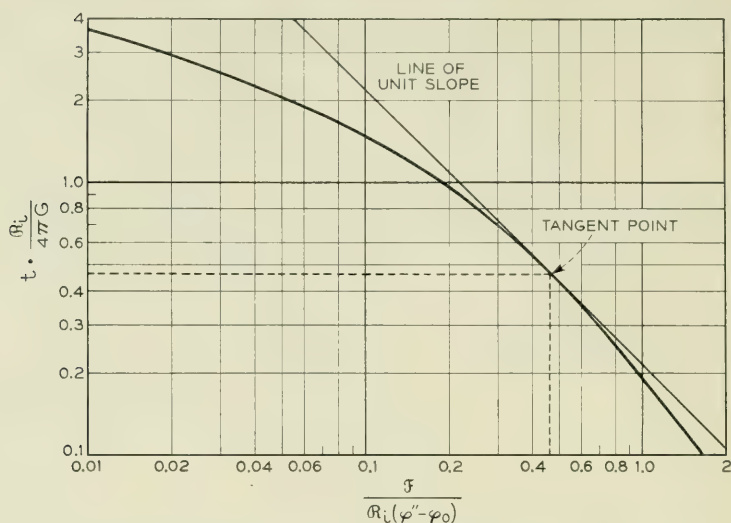


Fig. 6 — General relation between release time and ampere turns.

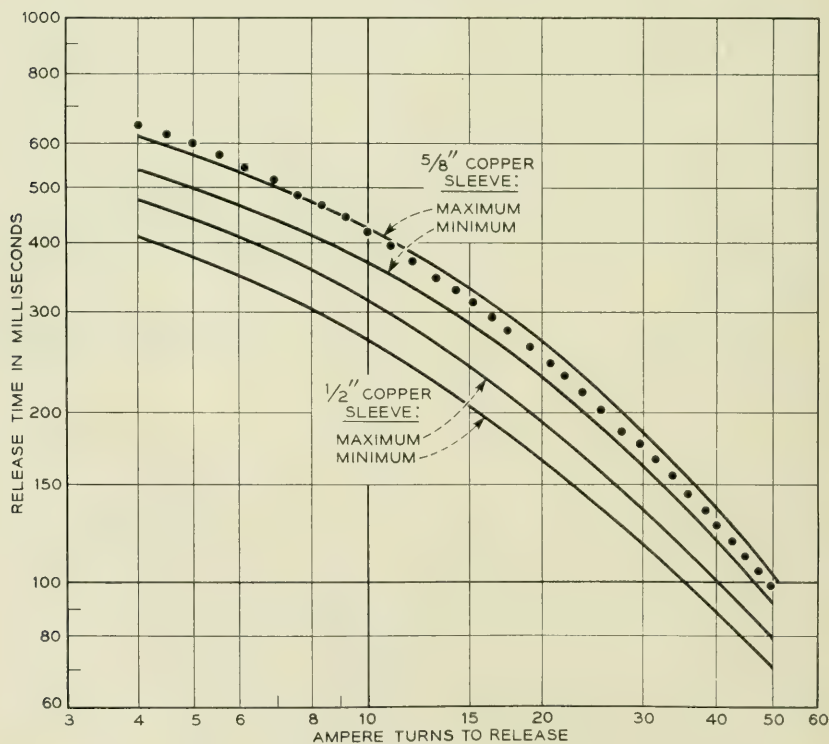


Fig. 7 — Observed release time characteristics.

release time is relatively independent of magnetic variations, provided it is adjusted to release at a specified ampere turn value.

The range in which t is inversely proportional to $\mathfrak{F}$ is that in which the logarithmic plot of Fig. 6 has a slope near unity. The point of tangency with a line of unit slope coincides with the value of z for which the bracketed function of z in (9) is a maximum. By equating the derivative of this to zero, it is found that this maximum occurs for $z = 3.09$, corresponding, from (8), to a value of 0.46 for $\mathfrak{F}'(\mathfrak{R}_i(\varphi'' - \varphi_0))$. From (6), the corresponding value of $t\mathfrak{R}_i'/(4\pi G)$ is 0.46. Thus by drawing a line of unit slope tangent to the observed relation between t and $\mathfrak{F}$, and reading the co-ordinates of the point of tangency, $4\pi G'(\mathfrak{R}_i$ and $\mathfrak{R}_i(\varphi'' - \varphi_0)$ may be evaluated. If G is known, these suffice to determine the values of $\mathfrak{R}_i$ and $\varphi'' - \varphi_0$.

Thus the release time is given by (6), and may be expressed in terms of the flux φ at which release occurs by means of (7), or in terms of the corresponding value of $\mathfrak{F}$, or $4\pi NI$, by means of (8). To relate the time to the spring load determining release, expressions are required relating the pull to φ or to $\mathfrak{F}$. To relate both time and pull characteristics to the design requires means for evaluating the magnetic constants and G in terms of the dimensions and materials of the design. The magnetic constants and the pull relations are discussed in the two following sections.

3 DECREASING MAGNETIZATION RELATIONS

To determine the decreasing magnetization relation experimentally, the magnet is demagnetized, and a measurement made of the flux developed on applying a full "soak," or high ampere turn value. The decreasing flux is then measured as the applied current is reduced, and finally reversed to determine $\mathfrak{F}_c$, the value required to restore the field to zero. The relation between φ and $\mathfrak{F}$ thus determined has the character shown in Fig. 1. If this curve conforms to the empirical relation given by (3) and (4), it is characterized by three constants: $\mathfrak{R}''$, φ'' , and either $\mathfrak{F}_c$, as in (3), or φ_0 , as in (4). These may be evaluated from measurements or estimated in preliminary design by the procedures indicated below.

Experimental Determination of Magnetic Constants

Equation (3) may be written in the form:

$$\frac{\mathfrak{F}_c + \mathfrak{F}}{\varphi} = \mathfrak{R}'' + \frac{\mathfrak{F}_c + \mathfrak{F}}{\varphi''}. \quad (10)$$

As $\mathfrak{F}_c$ may be read directly from the measured curve, as indicated in Fig. 1, values of $(\mathfrak{F}_c + \mathfrak{F})/\varphi$ may be computed from corresponding values of φ and $\mathfrak{F}$, and plotted against the corresponding value of $\mathfrak{F}_c + \mathfrak{F}$. One such plot for a slow release relay model is shown in Fig. 8. In agreement with (10), the plot is approximately linear over the range covered, and the slope and intercept may be used to evaluate $\mathcal{R}''$ and φ'' as indicated in the figure. If co-ordinate paper is used, as in Fig. 8, with radial lines spaced to give a convenient scale of φ , the value of φ'' is that corresponding to the radial line parallel to the plot. The observed linearity of this relation does not extend to much higher values of $\mathfrak{F}$, and the observed asymptote φ'' is fictitious, as the full curve is concave downwards, asymptotic to a lower value of φ than that found in this way. If the observed value of φ_0 does not agree with that computed from values of the other constants, it is preferable to use the observed rather than the computed value of φ_0 .

As shown in the preceding section, values of $\mathcal{R}_i$ (which is related to $\mathcal{R}''$ by (5)) and of $\varphi'' - \varphi_0$ may be independently determined from measurements of release time versus release ampere turns. In cases of disagreement, these values are to be preferred to those determined from the decreasing magnetization measurements. The latter are primarily of interest in checking design estimates, and in indicating the effects of variations in dimensions and material properties.

Estimation of Magnetic Constants

In the "tight" magnetic circuit of a slow release relay, the leakage field may be ignored, and the reluctance taken as the sum of the iron reluctance and that of the air gaps, which may be designated $\mathcal{R}_E$. The estimation of $\mathcal{R}_E$ is discussed in the following section, in connection with the pull relations. The iron reluctance is substantially that of a permanent magnet, supplying flux to the external circuit of reluctance $\mathcal{R}_E$. In using (3) to characterize the magnetization relation, the demagnetization in the second quadrant is taken as continuous with the decreasing magnetization in the first quadrant. Thus the value of $\mathfrak{F}_c$ is that which would apply to a permanent magnet of length ℓ , equal to that of the iron path, as given by:

$$\mathfrak{F}_c = H_c \ell, \quad (11)$$

where H_c is the coercive force of the material. For the soft magnetic materials used for slow release relays, H_c is of the order of 0.5 to 1.0 oersteds, so if $\ell = 10$ cm, $\mathfrak{F}_c$ lies in the range from 5 to 10 gilberts (4 to 8 ampere turns).

Similarly, the initial iron reluctance corresponds to the slope of the demagnetization curve at H_c , as given by the initial permeability μ'' . Thus the total initial reluctance α'' is given by the equation:

$$\alpha'' = \alpha_E + \frac{1}{\mu''} \sum \frac{\ell}{a}, \quad (12)$$

where $\sum \ell/a$ denotes the sum of terms corresponding to the component parts of the iron path, each term giving the length of the part divided by its cross sectional area. As μ'' is of the order of 20,000 for soft magnetic materials, the iron reluctance term is small compared with α_E , which is typically of the order of 0.010 cm^{-1} or less.

The value of φ'' may be estimated as the saturation flux for the core, or $B''a$, where a is the cross sectional area of the core. B'' is the saturation density, which may be taken as the value of B_M listed in a companion article.⁶ Estimates of φ'' thus obtained are lower than those which best fit the decreasing magnetization measurements. No convenient means for correcting this disparity has been found, other than an arbitrary increase by a factor of 20 per cent, based upon experience. In a particular relay design, however, the observed values of φ'' vary directly with the core cross section.

4 PULL RELATIONS

As shown by equation (6), the release time varies inversely as the incremental reluctance α_i . This is proportional to α'' , in which the dominant

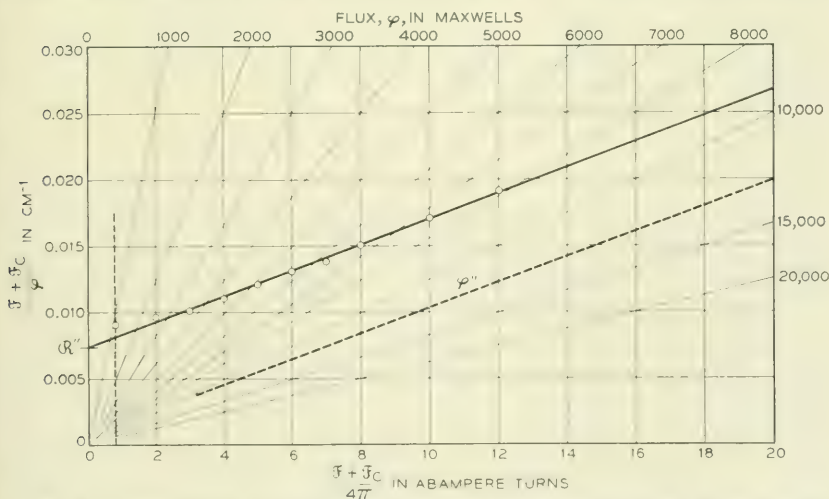


Fig. 8 — Evaluation of magnetic constants from measurements of decreasing magnetization.

term, from (12), is the total air gap reluctance $\mathcal{R}_E$. To obtain long delays, therefore, $\mathcal{R}_E$ must be made small, and consequently sensitive to small dimensional variations. These may be compensated in an initial adjustment, but subsequent changes must be minimized if constancy of performance is to be attained.

Two expedients have been used to provide a small and stable gap reluctance. The older one is the use of a "residual screw," an adjustable non-magnetic member which serves as an armature stop and assures a small air gap at the pole face. With this scheme, the residual screw is used to adjust the relay. The alternative scheme, used in the flat type relays of the Bell System, is to employ a domed pole face on the armature, providing a spherical surface in contact with a mating plane surface on the core. The only effective air gap at the point of contact is that of the chrome finish on the parts. With this scheme, the relay is adjusted by varying the spring tension, and thus the operated load.

General expressions for the pull of electromagnets are discussed in a companion article,⁶ where it is shown that the pull F provided by a gap flux φ is given by the equation:

$$F = \frac{\varphi^2}{8\pi} \frac{d\mathcal{R}_G}{dx}, \quad (13)$$

where x is the dimension in the direction of the pull, and $\mathcal{R}_G$ is the gap reluctance. In the usual case, $\mathcal{R}_G$ varies linearly with x , and $d\mathcal{R}_G/dx$ has the constant value $1/A$, where A is the effective pole face area. This is applicable to any case of plane mating surfaces having an appreciable separation, including the configuration usually employed with a residual screw as separator.

Pull for a Domed Pole Face

In the case of a domed pole face there is a concentration of the field near the point of contact, which varies with the effective air gap at this point. An expression for the reluctance can be developed for the idealized configuration shown in Fig. 9. In this, R is the radius of a spherical surface mating with a plane over the projected area A bounded by the radius αR . The separation x is that measured at the center of A . As indicated in the figure, an expression can be obtained for the gap reluctance $\mathcal{R}_G$ in terms of its reciprocal, or permeance. The latter is given by the integration of the permeances of the differential rings within the projected area. $\mathcal{R}_G$ is conveniently expressed in terms of the ratio $\mathcal{R}_\infty/\mathcal{R}_G$

given by the equation:

$$\frac{\mathfrak{R}_\infty}{\mathfrak{R}_G} = 2\pi R \mathfrak{R}_\infty \ln \left(1 + \frac{1}{2\pi R \mathfrak{R}_\infty} \right), \quad (14)$$

in which $\mathfrak{R}_\infty = x/A$, the value of $\mathfrak{R}_G$ which would apply if the spherical radius were infinite, and both mating areas plane. On substituting this expression for $\mathfrak{R}_G$ in (13), there is obtained the following equation for the pull at a domed pole face:

$$F = \frac{\varphi^2}{8\pi A} \left(\frac{\mathfrak{R}_G}{\mathfrak{R}_\infty} \right)^2 \frac{2\pi R \mathfrak{R}_\infty}{1 + 2\pi R \mathfrak{R}_\infty}. \quad (15)$$

The right hand side of (14), and the expression given by (15) for the ratio of F to $\varphi^2/(8\pi A)$ are both functions of a single dimensionless parameter: $R\mathfrak{R}_\infty$. These two ratios are plotted against this parameter in Fig. 10. These curves show how the reluctance and pull values compare

For $\theta < \alpha$ small,

$$r = R \sin \theta = R\theta$$

$$h = R(1 - \cos \theta) = \frac{R\theta^2}{2}$$

Gap Reluctance

$$\begin{aligned} \frac{1}{\mathfrak{R}_G} &= \int_0^{\alpha R} \frac{2\pi r \times dr}{x + h} \\ &= \int_0^{\alpha} \frac{2\pi R^2 \theta d\theta}{x + \frac{R\theta^2}{2}} \\ &= 2\pi R \ln \left(1 + \frac{R\alpha^2}{2x} \right) \end{aligned}$$

$$\text{Let: } \mathfrak{R}_\infty = \frac{x}{A} = \frac{x}{\pi(\alpha R)^2}$$

$$\frac{\mathfrak{R}_\infty}{\mathfrak{R}_G} = 2\pi R \mathfrak{R}_\infty \ln \left(1 + \frac{1}{2\pi R \mathfrak{R}_\infty} \right)$$

Pull for Flux φ

$$\begin{aligned} F &= \frac{\varphi^2}{8\pi} \times \frac{d\mathfrak{R}_G}{dx} = \frac{\varphi^2}{8\pi \cdot 1} \times \frac{d\mathfrak{R}_G}{d\mathfrak{R}_\infty} \\ &= \frac{\varphi^2}{8\pi \cdot 1} \times \left(\frac{\mathfrak{R}_G}{\mathfrak{R}_\infty} \right)^2 \times \frac{2\pi R \mathfrak{R}_\infty}{1 + 2\pi R \mathfrak{R}_\infty} \end{aligned}$$

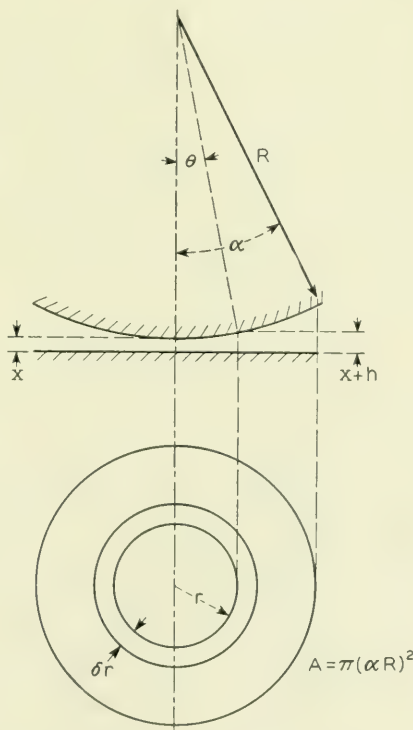


Fig. 9 — Reluctance and pull of a domed pole face.

with those for a gap of the same separation and projected area, but with plane pole faces (a dome of infinite radius).

Ampere Turn Sensitivity

The effect of the relation between F and φ on the release time cannot be conveniently deduced from (6), which involves both the term in z , which varies with φ , and the reluctance $\mathcal{R}_i$, which also enters the pull relation. If the pull is expressed in terms of $\mathfrak{F}$, however, (9) shows that the time varies inversely with $\mathfrak{F}$, and is substantially independent of the reluctance, provided z is in the usual range from 2 to 6. It follows that maximum release time can be obtained for a given load if the value of $\mathfrak{F}$

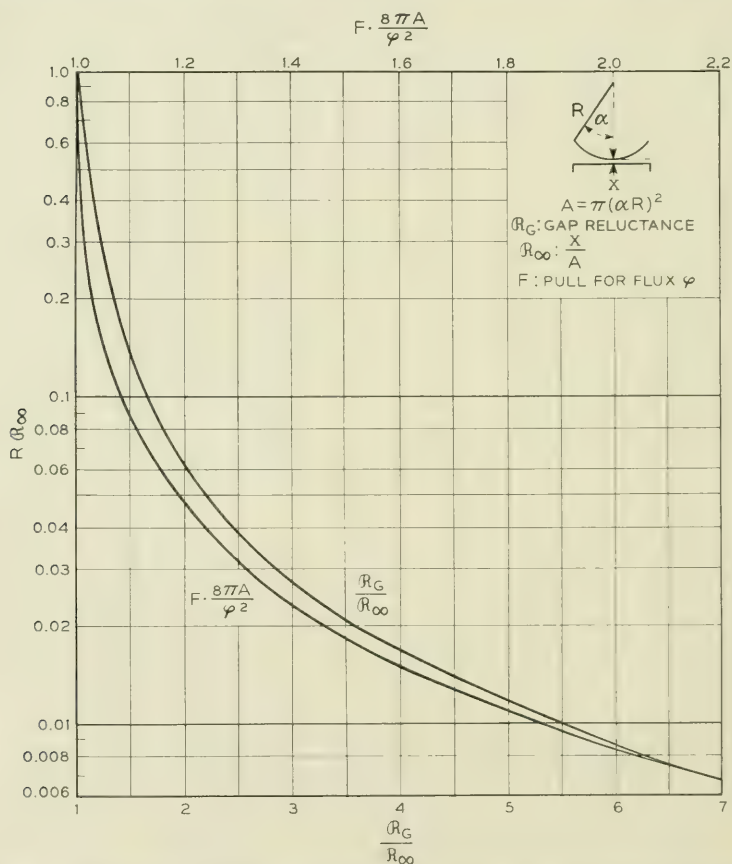


Fig. 10 — Reluctance and pull relations of a domed pole face electromagnet.

required to operate this load is minimized. Thus maximum release time is attained by providing maximum ampere turn sensitivity.

A general expression for the relation between F and $\bar{\Phi}$ may be obtained by substituting in (13) the expression for $(\bar{\Phi}_c + \bar{\Phi})/\varphi$ given by (10). A much simpler relation is given by the linear approximation to the decreasing magnetization curve in which $(\bar{\Phi}_c + \bar{\Phi})/\varphi$ is taken as equal to $\mathfrak{R}_i$. This approximation has the same slope at φ_0 as the curve given by (3) or (10). The release pull usually corresponds to values of φ much nearer φ_0 than φ'' , and the approximation is satisfactory in this range if φ_0/φ'' is small. Writing $(\bar{\Phi}_c + \bar{\Phi})/\mathfrak{R}_i$ for φ in (13), there is obtained the equation:

$$F = \frac{2\pi(NI + (NI)_c)^2}{\mathfrak{R}_i^2} \frac{d\mathfrak{R}_G}{dx}, \quad (16)$$

in which $(NI)_c$ is written for $\bar{\Phi}_c/(4\pi)$. This approximation is used for convenience and simplicity. The general expression, which is required for higher values of $\bar{\Phi}$, is obtained by substituting the right hand side of (10) for $\mathfrak{R}_i$ in (16) and in the expressions derived from it.

By comparison with (12), $\mathfrak{R}_i$ may be taken as equal to the gap and joint reluctance $\mathfrak{R}_E$ plus a modified and minor term for the iron reluctance. For the present purpose, it is convenient to take $\mathfrak{R}_i$ as the sum of the main gap reluctance $\mathfrak{R}_G$, which varies with x , and $\mathfrak{R}_F$, which includes the iron reluctance and the constant reluctances of the heel gap and of any joints in the magnetic structure. For a plane pole face, $\mathfrak{R}_G$ is given by x/A , where A is the effective pole face area, and (16) reduces to the equation:

$$F = \frac{2\pi(NI + (NI)_c)^2}{A \left(\mathfrak{R}_F + \frac{x}{A} \right)^2}.$$

As shown in one of the companion articles,⁶ the pull F at travel x for a given value of NI is a maximum when the gap reluctance x/A is equal to the reluctance $\mathfrak{R}_F$ external to the gap. A similar relation applies for a domed pole face gap, as can be shown from the expressions given above. Taking the linear approximation to apply, $(\bar{\Phi}_c + \bar{\Phi})/(\mathfrak{R}_F + \mathfrak{R}_G)$ may be substituted for φ in (15). Writing $x/\mathfrak{R}_F$ for A in the resulting equation, this may be written in the form:

$$F = \frac{2\pi(NI + (NI)_c)^2}{x\mathfrak{R}_F} \Omega,$$

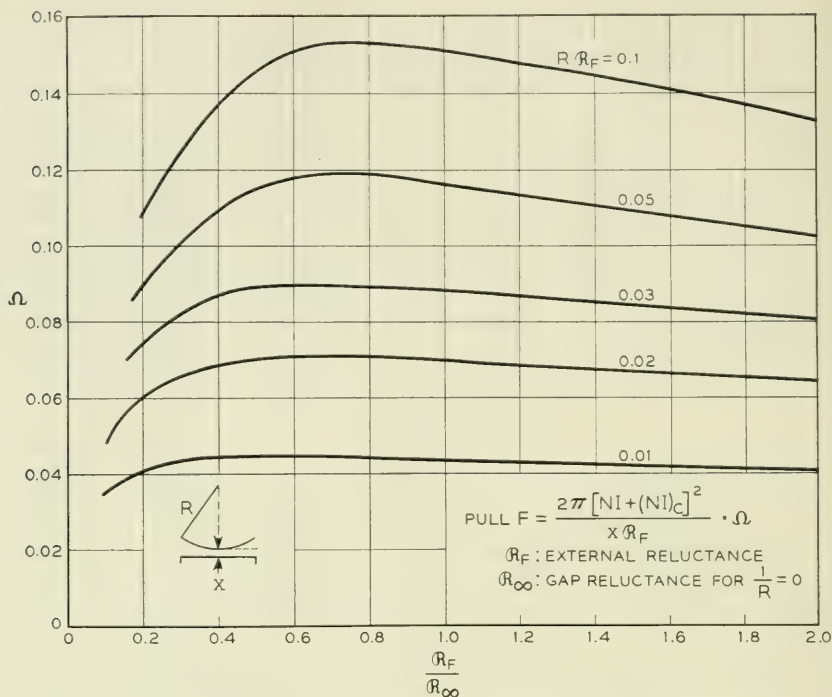


Fig. 11 — Ampere turn sensitivity of a domed pole face electromagnet.

where:

(17)

$$\Omega = \frac{R_G^2 R_F}{R_\infty (R_G + R_F)^2} \frac{2\pi R R_\infty}{1 + 2\pi R R_\infty}.$$

The ratio Ω is a function of R_F/R_∞ and of $R R_\infty$, which determines R_G/R_∞ , as shown by (14). Alternatively, Ω may be taken as a function of the ratios R_F/R_∞ and $R R_F$, and Ω may be represented, as in Fig. 11, by a family of curves giving Ω versus R_F/R_∞ for various values of $R R_F$.

From (17), maximum ampere turn sensitivity is attained by minimizing the separation x and the reluctance R_F external to the gap. Assuming these to be made as small as engineering considerations permit, the pull for a given value of NI varies as Ω . With x and R_F fixed, Ω now depends only on the dome radius R , to which $R R_F$ is now proportional, and on the projected area A , to which R_F/R_∞ is now proportional. From the curves of Fig. 11 it is apparent that maximum ampere turn sensitivity is attained by using as large a value of dome radius as possible. For a given dome radius, there is an optimum value of R_F/R_∞ ,

and hence an optimum value of A , corresponding to the maximum shown by each curve in Fig. 11.

AG Relay Pull

The *AG* relay³ is a slow release relay with a domed pole face, to which the relations given above apply approximately. M. A. Logan and O. C. Worley developed a more exact expression for the pull in this case, in which an increment to the pull is given by the secondary pole faces on the side legs of the armature, which mate with the side legs of the *E* shaped core member. A further increment to the pull is given by the area at the main gap which lies outside the dome proper. Their analysis may be summarized in the notation used here by writing:

$$\mathfrak{R}_i - \mathfrak{R}_I = \mathfrak{R}_S + \frac{\mathfrak{R}_G \mathfrak{R}_P}{\mathfrak{R}_G + \mathfrak{R}_P},$$

where R_I is the iron reluctance, $\mathfrak{R}_S$ is the side gap reluctance, and the main gap reluctance is that of the domed gap $\mathfrak{R}_G$ in parallel with that of the remaining area, $\mathfrak{R}_P$. Substituting $R_i - \mathfrak{R}_I$ for $\mathfrak{R}_G$ in (16), there is obtained the following expression for the pull:

$$F = \frac{2\pi(NI + (NI)_c)^2}{\mathfrak{R}_i^2} \left(\frac{d\mathfrak{R}_S}{dx} + \left(\frac{\mathfrak{R}_G}{\mathfrak{R}_G + \mathfrak{R}_P} \right)^2 \frac{d\mathfrak{R}_P}{dx} + \left(\frac{\mathfrak{R}_P}{\mathfrak{R}_G + \mathfrak{R}_P} \right)^2 \frac{d\mathfrak{R}_G}{dx} \right). \quad (18)$$

An expression for $d\mathfrak{R}_G/dx$ is given in Fig. 9, while $d\mathfrak{R}_S/dx$ and $d\mathfrak{R}_P/dx$ are given by $1/A_S$ and $1/A_P$, where A_S and A_P are the effective pole face areas for these two gaps. For the dimensions applying to the *AG* relay, these additional terms introduce minor but significant corrections to the values of F computed from (17).

Engineering Pull Data

In determining the requirements for slow release relays, the estimation of the range of variation in pull is a major problem. A procedure for such estimation developed by M. A. Logan makes effective use of the relation between F and $(NI + (NI)_c)$ in (16). This relation gives a linear plot of slope two when F is plotted against $(NI + (NI)_c)$ on logarithmic paper. A basic plot of this nature may be experimentally determined for a model having nominal values of coercive force H_c , to which $(NI)_c$ is proportional, and of finish thickness and side gap separa-

tion. The effect of known changes in these two latter factors may be determined by the observed vertical shift in this logarithmic plot, corresponding to changes in $\mathcal{R}_i$ and $d\mathcal{R}_G/dx$. The effect of a change in H_c is a change in the value of $(NI)_c$ to be subtracted from $(NI + (NI)_c)$ in determining NI . From these results there could be prepared curves giving the relation between F and NI for different combinations of the several variables for which allowance should be made.

5 EVALUATION OF CONDUCTANCE

The preceding sections have described procedures for estimating and measuring all the terms entering the expressions for the release time except the conductance G . This quantity is the sum of the sleeve conductance G_s and the eddy current conductance G_E . In some cases a short circuited winding may be used instead of a sleeve; in such cases G is the sum of G_E and the coil constant of the winding, G_c or N^2/R .

Sleeve Conductance

The conductance of a cylindrical sleeve is the sum of the conductances of the differential shells of length ℓ , radius r and thickness δr , as indi-

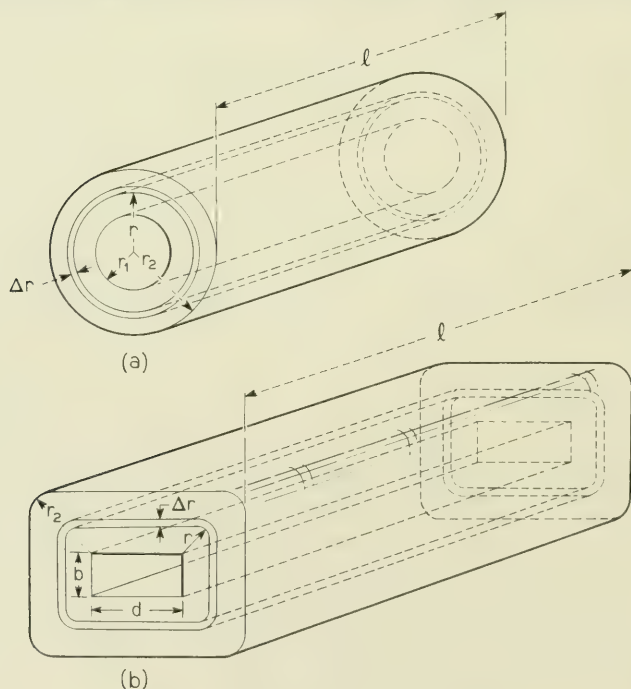


Fig. 12 — Sleeve conductance relations.

cated in Fig. 12(a). Thus the total conductance G_s is given by:

$$G_s = \int_{r_1}^{r_2} \frac{\ell dr}{2\pi\rho r},$$

where ρ is the resistivity of the material. On integration, there is obtained:

$$G_s = \frac{\ell}{2\pi\rho} \ln \frac{r_2}{r_1}. \quad (19)$$

The value of ρ for copper is 1.73×10^{-6} ohm-cm. If r_2 is twice r_1 , for example, and $\ell = 5$ cm, the value of G_s given by (19) for a copper sleeve is 320,000 mhos.

In the case of a sleeve of rectangular section, as shown in Fig. 12(b), an approximation may be obtained by taking the sleeve as made up of shells with straight sides parallel to the center hole, connected by quarter circles. Then the perimeter of the shell at a distance r from the center hole is $2(b + d + \pi r)$. The total sleeve conductance is therefore given by:

$$G_s = \int_0^{r_2} \frac{\ell dr}{2(b + d + \pi r)\rho},$$

in which r_2 is the wall thickness. On integration, there is obtained the equation:

$$G_s = \frac{\ell}{2\pi\rho} \ln \left(\frac{b + d + \pi r_2}{b + d} \right). \quad (20)$$

This expression is identical with that for the cylindrical sleeve, as given by (19), when the ratio r_2/r_1 of the radii is equal to the ratio of $(b + d + \pi r_2)$ to $(b + d)$.

Coil Conductance

For a cylindrical coil, the number of turns N is determined by the area of the coil section cut by a plane through the axis, or $\ell(r_2 - r_1)$, where ℓ is the length of the coil and r_1 and r_2 are its inner and outer radii. If a is the cross sectional area of the wire, and c the fraction of the coil space occupied by the conductor,

$$Na = c\ell(r_2 - r_1).$$

The mean length of turn is $\pi(r_2 + r_1)$, and the total length of conductor is N times this. Hence the resistance R is given by:

$$R = \frac{\rho N \pi (r_2 + r_1)}{a}.$$

The expression for N/R given by these equations gives the following expression for the coil constant N^2/R , or G_c :

$$G_c = \frac{e\ell}{\pi\rho} \frac{r_2 - r_1}{r_2 + r_1}. \quad (21)$$

For unit length of coil, values of G_c are shown in Fig. 13, plotted against r_2/r_1 for various values of e , together with the corresponding relation for G_s , as given by (19). This shows that the sleeve provides the maximum value of conductance for the space occupied. It also shows that the space used for a given value of sleeve conductance reduces the value of G_c that can be provided. When the full depth of the winding space is used by both coil and sleeve, called a slug in this case, each occupying part of the length, the value of G_c attainable is reduced in proportion to the length used by the slug. When the coil is outside the sleeve, the outer radius of the sleeve is the inner radius of the coil, and the value of G_c is reduced in proportion to the value of G_s . For a given value of e (given wire size and insulation) the value of G_c is fixed by the winding space and the value of G_s .

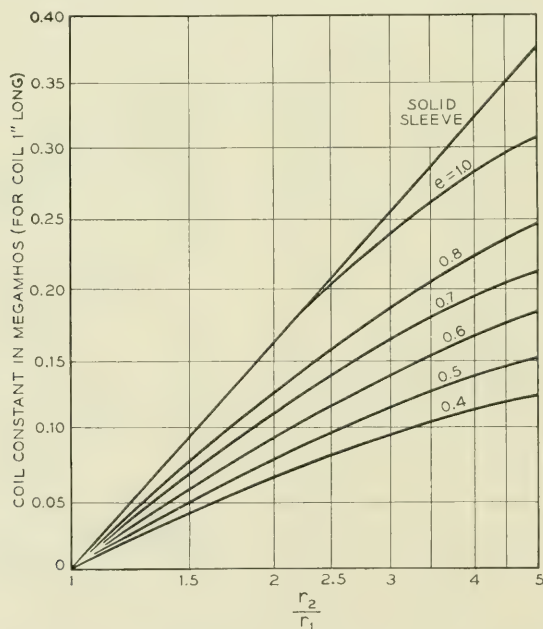


Fig. 13 — Relation between coil constant and coil dimensions.

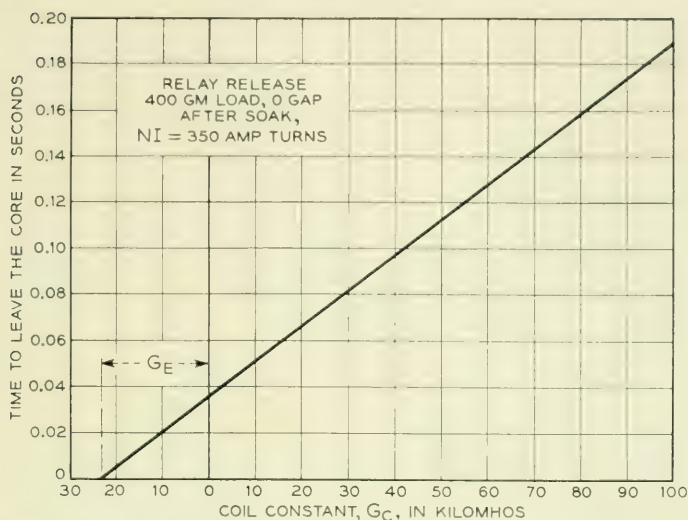


Fig. 14 — Experimental evaluation of eddy current conductance for release of relay.

Eddy Current Conductance

In one of the companion articles⁴ it is shown that when the eddy current conductance G_E is a minor term in G , as with slow release relays, it is given by the equation:

$$G_E = \frac{\ell}{8\pi\rho}, \quad (22)$$

where ρ is the resistivity of the material, and ℓ is the length of the magnetic path. For iron, $\rho = 11 \times 10^{-6}$ ohm-cm, so for $\ell = 5$ cm, the value of G_E given by (21) is 17×10^3 mhos. The equation applies to a path of uniform circular cross section, so that the effective value of ℓ for most relay structures is intermediate between that of the core and that for the complete path.

The eddy current conductance of a specific model may be experimentally determined by measuring the release time with a winding shorted through an external resistance. A series of measurements are made in which this resistance is varied, while the initial current, which determines the "soak NI " value, is kept constant. The different values of the resistance correspond to different values of the coil constant G_C or N^2/R . From (6), the time t in this series of measurements varies directly as G , or $G_C + G_E$. Then a plot of t versus G_C , as illustrated in Fig. 14, is linear, and has a negative intercept giving the value of G_E .

Values of G_E thus determined agree with those estimated from (22) within the level of uncertainty as to the applicable value of ℓ in this equation.

6 DESIGN OF SLOW RELEASE RELAYS

The following discussion is confined to the bearing on design decisions of the performance relations developed in the preceding sections, and does not cover the manufacturing considerations involved. The development of a specific design depends on the initial choice between certain alternative features which require description.

Design Alternatives

The features considered here are: (1) the adjustment means, (2) the form of sleeve, (3) the criterion of adjustment.

The two methods of adjustment that have been used are (a) residual screw adjustment, (b) spring load adjustment. The former affects the release time by changing the reluctance and the residual flux, the latter by changing the flux or ampere turn value at which release occurs. The differences in the two methods relate more to the stability of the adjustment than to its ease or initial accuracy. Spring adjustment permits the use of the domed pole face, which is inherently stable except as the finish thickness may be affected by wear.

The two forms of sleeve are the interior sleeve, which uses the full length of the winding space, but only part of the depth, and the slug, which uses the full depth, but only part of the length. As shown in Section 5, the values of G_s and G_c , or coil constant N^2/R , attainable with a given winding space are independent of which arrangement is used. The slug provides some flexibility as to operate time, on which its retarding effect is a maximum when it is near the gap, and a minimum when it is away from it. It is subject to smaller temperature changes, with consequent changes in conductance, than the sleeve. The other differences between the two arrangements relate to manufacture, and to the costs of both coil and sleeve. The interior sleeve is used with the flat type relays of the Bell System.

The two different criteria of adjustment used are the release current, and the release time. The former is an indirect control, using a measurement of release ampere turns to determine the time that will result for the sleeve conductance used; the other is a direct measurement of the quantity to be controlled. In principle, the latter method would appear preferable, but its use is attended with several disadvantages.

The most important of these is the uncertainty as to the sleeve temperature that applies to any measurement made in central office maintenance, unless the relay to be tested is cut out of service for an hour or more before measurement. The conductivity of copper varies approximately as its absolute temperature, or, for engineering estimates, as $390 + T_F$, where T_F is the temperature in degrees Fahrenheit. Coil temperatures of 225°F are permitted in normal relay operation. As the time varies as the sleeve conductance, a relay with its sleeve at this temperature would have a release time in the ratio 470/615 or 76/100 to the rated release time for 80°F. Allowance for variation in this range is made in circuit design, but a corresponding uncertainty as to the condition applying in adjustment would effectively double this variation. Current flow adjustment is free of this difficulty, and the variations in the correlation of release time with release ampere turns are less than those resulting from the temperature uncertainty in any convenient procedure for timing measurements. Current flow adjustment has the further advantage of using equipment that is employed for other relays in central office maintenance. It is the more commonly used criterion of adjustment for Bell System relays.

Operate Considerations

A slow release relay must not only provide the desired release performance; it must also operate its load. The operate pull characteristics are similar to those of other relays of the same general type, as the domed pole face, in particular, gives nearly the same pull at an open gap as a plane pole face of the same total area. Thus the pull characteristics of the *AG* relay are similar to those of the *AJ* relay³ for the same travel. The sleeve retards the flux development, and makes operation slower than that for the same coil input without the sleeve. In most applications of slow release relays, this has little or no effect on circuit operation. Faster operation can be obtained by increasing the steady state power applied, but this is limited by heating considerations. The large part of the winding space used for the sleeve limits the operate sensitivity of slow release relays, and increases the power required for a given load. The load for a given relay design determines a minimum ampere turn value for operation, and this is related to the steady state power by the identity: $(NI)^2 = I^2 R N^2/R$. As the coil constant N^2/R , or G_c , is determined by the available winding space available for the coil, the power requirements of slow release relays are higher than those of similar relays having the full winding space available for the coil.

Optimum Design Conditions

In considering the design features that are favorable to slow release operation it is convenient to refer again to the expression for the release time given by equation (9):

$$t = \frac{4\pi G(\varphi'' - \varphi_0)}{4\pi NI} f(z), \quad (9)$$

where $f(z)$ is the function shown in Fig. 5 to be nearly a constant in the range of interest. It is also convenient to refer to the expression for the pull given by the equation (16):

$$F = \frac{2\pi(NI + (NI)_c)^2}{\mathcal{R}_i^2} \frac{d\mathcal{R}_G}{dx}.$$

Together with reference to the operate requirements, these two equations indicate the characteristics that are important for slow release operation.

The winding space determines the value of G attainable, as shown by the relations of Section 5. It limits the combined values of G_s and G_c , of which the latter controls the operate power sensitivity, while the former, from (9), determines the release time. Thus both the operate sensitivity and the release time attainable vary directly with the winding space and hence with the over-all size of the relay.

From (9), the attainable release time is nearly proportional to φ'' , and hence to the cross section of the core, assuming φ_0/φ'' to be small. The external core dimension is the internal dimension of the sleeve, so that increases in φ'' are offset by decreases in G if this dimension alone is varied. In any case, sufficient section must be provided for φ'' to have a margin over the field required to operate the maximum load.

In telephone use, the slow release relays are a minority group in a relay population which must, for maximum economy in manufacture and use, have common overall dimensions and as few differences as are consistent with the requirements of specific uses. Thus the core section, winding space, and over-all dimensions reflect an optimum choice for the whole relay population, and not for the slow release relays alone. The latter are distinguished by as few special features as are essential to their special function. In the AG relay these are: the armature, the heat treatment of the magnetic parts, the sleeve and coil, and a buffer spring for load adjustment.

The material and its heat treatment determine the iron reluctance and the coercive mmf, $4\pi(NI)_c$, of which the latter is the more important quantity. It enters the timing relation indirectly in the minor term φ_0 ,

which varies directly with $(NI)_c$, and enters the pull relation (16) directly as a major term. If high coercive material were used to increase $(NI)_c$ and thus reduce the release ampere turns NI , the effect of the latter on the time would be offset by the increase in φ_0 . The latter varies also with the reluctance $\mathcal{R}_i$, and if φ_0/φ'' is not small, the time for a given value of NI is no longer independent of the reluctance variation, and the release ampere turn value is unsatisfactory as a criterion of adjustment. Thus a small stable value of $(NI)_c$, as given by low coercive material, is preferable when the release current is used as the criterion of adjustment.

The pole face dimensions are determined by the need for a low stable value of gap reluctance. As shown in Section 4, maximum release sensitivity is obtained by using as large a dome radius as will assure a consistent configuration, together with a projected area optimum with respect to the reluctance in series with the gap. Fig. 11 shows this optimum to be broad, allowing considerable latitude in the choice of the projected area with reference to other design considerations.

The finish applied to the magnetic parts at the pole face is critical with respect to wear and stability, as well as to the danger of mechanical adhesion which would affect the release load. A nickel chrome finish is used in the *Y* and *AG* relays. The effective air gap thus introduced is the dimension x of Figs. 9 and 10. As these figures show, this has a relatively large effect on the pull relation, and any change in this dimension as a result of wear tends to increase the pull and the release time. The analysis of Section 4 shows that maximum sensitivity is attained with as small a finish separation as can be provided, which is fortunately in agreement with the manufacturing considerations for the finish.

The adjustment means used with a fixed pole face, e.g., the dome type, is some form of spring adjustment for controlling the operated load. The normal operated load of a relay is the product of the contact force and the number of contacts, plus the back tension and the force required for spring flexure. Adjustment of the back tension is used, but is limited by the requirements for operation. Hence a buffer spring is employed in *AG* relays, which is picked up in the last few mils of armature travel and affords control of the operated load. This spring is adjusted by bending, and a tolerance range of 50 gm wt is allowed in this adjustment.

Performance Variations

Performance variations fall into two categories: the initial differences among individual relays that are nominally identical, and the further differences that develop in service as a result of wear and other changes.

In the latter category, the changes may be magnetic or mechanical. The mechanical changes appear in the value of the operated load, and are practically negligible in relays of the AG type, because of the lift-off actuation used.³ In this they compare favorably with Y-type relays, which are liable to load changes through stud and contact wear. The magnetic changes include "ageing" of the magnetic material, a gradual change in permeability and coercive force which occurs with magnetic iron. Hydrogen annealed silicon steel, used in the AG relay, is free from this defect. The other magnetic changes are those appearing in the pole face configuration and finish thickness as a result of wear. No significant changes of this type have been observed in the AG relay, which is therefore highly stable in maintaining the performance of its initial adjustment.

In use, the release time varies inversely with the sleeve or slug temperature. As shown above, the 140°F temperature rise allowed for coils in normal operation will, if experienced by the sleeve, reduce the release time by about 25 per cent. Allowance for variation in this range must be made in planning the applications of slow release relays. Relays fitted with slugs are less affected by coil temperatures than are those using interior sleeves. The slug arrangement, with an insulating air gap between slug and coil, would be essential to a slow release relay designed for closely controlled release time.

The initial differences from nominal performance may be sub-divided into those arising from the adjustment tolerance, and those resulting from using release current as the adjustment criterion. Only the former would apply if the direct timing criterion were used, provided a procedure was employed in which the sleeve temperature was fixed in measurement. A tolerance range of 50 gm wt is a 25 per cent variation for a representative load of 200 gm wt, and corresponds, from (16) to about a 12 per cent variation in $NI + (NI)_c$. If $(NI)_c$ is about equal to NI , the variation in the latter is about 25 per cent, which results, from (9) in a similar variation in the release time. A smaller tolerance range could be attained by more expensive adjustment means than the buffer spring of the AG relay, with a corresponding reduction in the range of time variation.

With release current as the criterion, the value of NI in (9) is subject to the adjustment variation, of the order of 25 per cent for the case cited above. In addition, the release time is subject to variations in G , $\varphi'' - \varphi_0$, and $f(z)$. In the optimum range of operation where $f(z)$ is nearly constant, variations in this term are minor. φ_0 is subject to the relatively large percentage variations in $(NI)_c$ and $\mathcal{R}_i$, but the net effect

of these is small if φ_0/φ'' is as small as 10 per cent, as in the *AG* relay. Hence the principle variations in the release time are those associated with the variations in the sleeve dimensions and the core cross section, which determine the values of G and φ'' respectively. They cause variations in the *AG* relay of from 20 to 30 per cent. Combined with the variation in N resulting from the adjustment tolerance, they result in release time variations of 40 or 50 per cent.

For the great majority of slow release relay applications, much larger variations have no effect on circuit performance, and for only a small minority is it desirable to employ requirements assuring this measure of control. It is apparent from the preceding discussion that changes in construction and adjustment procedure might be made which would result in closer time control: as they would add to the costs of both manufacture and use, they are not economically justified.

Engineering Data

A specific relay design is termed a code: all relays of the same code are nominally identical. A code is distinguished from other codes of the same relay type by the contacts and coil provided, and by the sleeve or other special features used. To design slow release relay codes, and to specify their adjustment requirements and the performance provided, requires means for determining from the general pull and timing characteristics the values applying to specific cases. For this purpose the pull and timing data are prepared in the form illustrated in Fig. 15.

The pull information appears in this chart in the form of the two curves marked "hold pull" and "release pull." These show the relation between operated load and release ampere turns for the two limiting cases of minimum and maximum pull, and reflect the variations in the terms of (16). No individual relay releases on ampere turns above the hold values, all relays release on ampere turns below the release values.

The timing information appears in the form of a pair of curves for each sleeve size that is used: one giving the minimum time, and one the maximum. Only one such pair is included in Fig. 15. The difference between the two curves represents the 20 to 30 per cent variation associated with the variations in G and $\varphi'' - \varphi_0$.

The information on a chart of this type is used in conjunction with engineering data for the operated loads of the various spring combinations, and the changes in these loads that can be provided by back tension and buffer spring adjustment.

For each spring combination, there is a lower limit to the adjusted

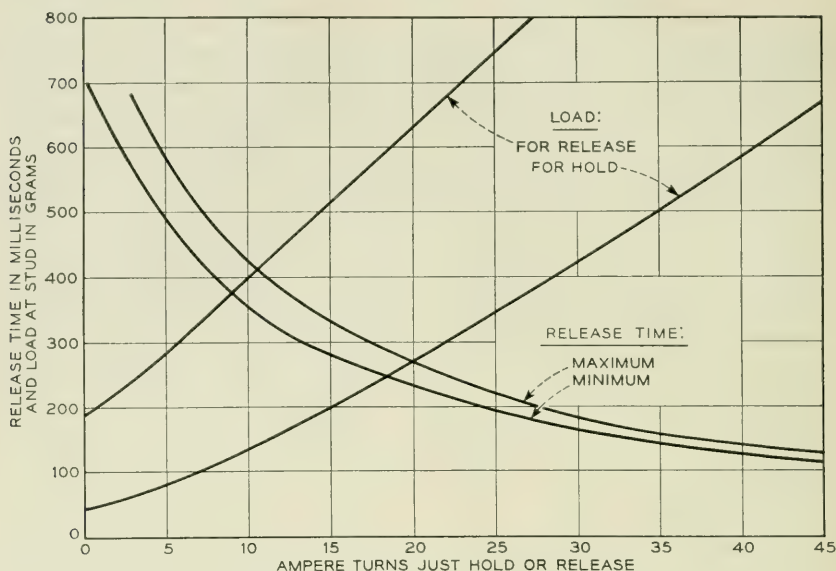


Fig. 15 — Engineering data for release time estimation.

load attainable in all cases. This corresponds to the case in which the contact force, which is not subject to adjustment, has its maximum value. The NI value read from the hold curve for this load is the lowest that can be specified as a "hold" requirement. Each individual relay has a release F versus NI characteristic intermediate between the release and hold capability curves. To meet the hold requirement its load must be adjusted, and this adjustment is subject to the tolerance cited above. The minimum load is set by the hold requirement, and the maximum load is set by a release requirement of a lower NI value, the difference between the two NI requirements corresponding to the tolerance range in load adjustment.

Thus adjustment values are determined in the form of limits to the ampere turn value at which release occurs: the lower limit is the release value, the upper the hold. The release time limits can then be read from the timing curves. The maximum time is read from the maximum curve at the release ampere turn value. The minimum time is read from the minimum curve at the hold ampere turn value. This minimum time, when determined for the largest sleeve, is the longest time that can be guaranteed for the load in question, and is subject to reduction in service by the temperature variation previously discussed. When a shorter release time is desired, a smaller sleeve may be used, or a higher hold

value specified, subject to the capacity of the adjustment springs to supply the necessary increase in load.

In the common case for which only the minimum time is of circuit importance, the release value is chosen without reference to the hold value, solely for the purpose of assuring that the relay will not lock up indefinitely. This procedure takes advantage of the simpler requirements of this case by widening the adjustment tolerance and reducing adjustment effort.

7 CONCLUSIONS

The relation between the release time of slow release relays and the design parameters can be more accurately expressed in analytical form than the other time characteristics of relays. These analytical relations, as presented in this article, can be used for the estimation of release time, and particularly for the determination of the effect on this time of variations in the design parameters. The need for a low reluctance magnetic circuit makes the performance of slow release relays highly sensitive to dimensional and material variations, and adjustment is required to assure the timing limits required in their use. Such use usually permits a wide spread in release time, provided a minimum value is assured. Advantage is taken of this in providing slow release relays which perform their function at a minimum cost in manufacture and use, materially lower than that for the construction and adjustment practices which would be required for closer timing control.

ACKNOWLEDGEMENTS

The specific references made to individuals and prior studies do not include all the work which has been drawn on in the preparation of this article. In particular, much of the discussion of the applications of the analysis is based on work carried out by M. A. Logan, Mrs. K. R. Randall, O. C. Worley and others in the development of the AG relay.

REFERENCES

1. F. A. Zupa, The Y-type Relay, Bell Lab. Record, **16**, p. 310, May, 1938.
2. H. N. Wagar, The U-type Relay, Bell Lab. Record, **16**, p. 300, May, 1938.
3. A. C. Keller, A New General Purpose Relay for Telephone Switching Systems, B.S.T.J. **31**, p. 1023, Nov., 1952.
4. R. L. Peek, Jr., and M. A. Logan, Estimation and Control of Operate Time of Relays, pages 109 and 144 of this issue.
5. H. N. Wagar, Slow Acting Relays, Bell Lab. Record, **26**, p. 161, April, 1948.
6. R. L. Peek, Jr., and H. N. Wagar, Magnetic Design of Relays, page 23 of this issue.

Economics of Telephone Relay Applications

By H. N. WAGAR

(Manuscript received July 6, 1953)

Today's telephone central office is largely built around the telephone relay. As each office may use some 60,000 of them, their performance characteristics, as well as their first cost, have a very important influence on the cost of the office. By properly balancing such factors economically, the lowest cost office can be planned.

This paper shows how performance features and design factors may all be expressed as "equivalent first cost," and so be related to manufacturing cost itself. The influence of lot-size on manufacturing cost is considered including a determination of optimum lot-sizes, to aid the relay designer in deciding on standardized models.

With a basic performance language formulated — equivalent first cost — optimum relations among the variables may be established. Such results are given for (a) optimum coil-plus-power costs, (b) optimum power-plus-speed costs, (c) optimum number of relay codes, and other related problems. Methods are also given to evaluate how serious the effect may be when optimum conditions are not satisfied.

Were it not for the application of these methods, central office costs would be much higher.

INTRODUCTION

The telephone central office of today is in part an enormous computing machine which, upon instructions from the customer, given by his dial, calculates how to find the called party; then the remaining mechanism completes the call. This machine is composed largely of interconnected relays, used in enormous quantities. Every dial call in a large city involves around 1000 relays. A typical large office contains more than 60,000 of them; in fact, they are used so extensively that the Western Electric Company manufactures about ten of them for every new subscriber, and their output is figured in tens of millions per year. It is not

surprising that relay use has an enormous influence on the cost of the central office — not just because of relay purchase cost but also the relay's influence on other office factors. For example, the size of the power plant and the total number of equipments for common control, which depend on the functioning time of the relays, are decided by the relay characteristics. In the application of relays, then, it may well prove that the largest economies can be realized by spending a little more for each relay in the beginning in order to save still more in the cost of the power plant, common control equipment, size of the building, and so forth. It has been found that attention to ways of optimizing the application costs for relays can lead to an appreciably lower cost central office, and this paper will illustrate a few cases of how such a problem is approached. Because the telephone system is so large, the influence of each relay, taken in total, is also large. Fortunately, at Bell Telephone Laboratories, it has been possible to take the over-all view of the subject through familiarity with all phases of the relay application problem; and large economies which will eventually benefit the customer are resulting.

The basic problem to be discussed is how best to realize maximum economy of the central office so far as relays and their uses are concerned. As in most engineering problems, it is necessary to evaluate and compromise between oppositely varying cost effects of such things as efficiency, manufacturing cost, amount of equipment, ease of mounting and wiring, maintenance in all its aspects, and the like. The end result of each of these variables is its economic effect on the office as a whole, and they can only be compared if their values can be stated on a comparable basis. It has been found that each such effect may be considered as an incremental cost over and above a reference cost for its particular ideal condition. If properly chosen so as to be independent, then all such incremental costs may be added. They then represent the net "cost penalty," compared to the design with all ideal conditions taken together, and describe the merit of the design. They can also be used to find the optimum design. The methods apply generally to many other similar problems.

BREAKDOWN OF THE PROBLEM

Consider first how relays are applied in the telephone switching system. To the greatest extent possible a basic relay structure is chosen, carefully planned for low maintenance effort, all of whose basic parts can be made by mass production methods. On this basic framework one

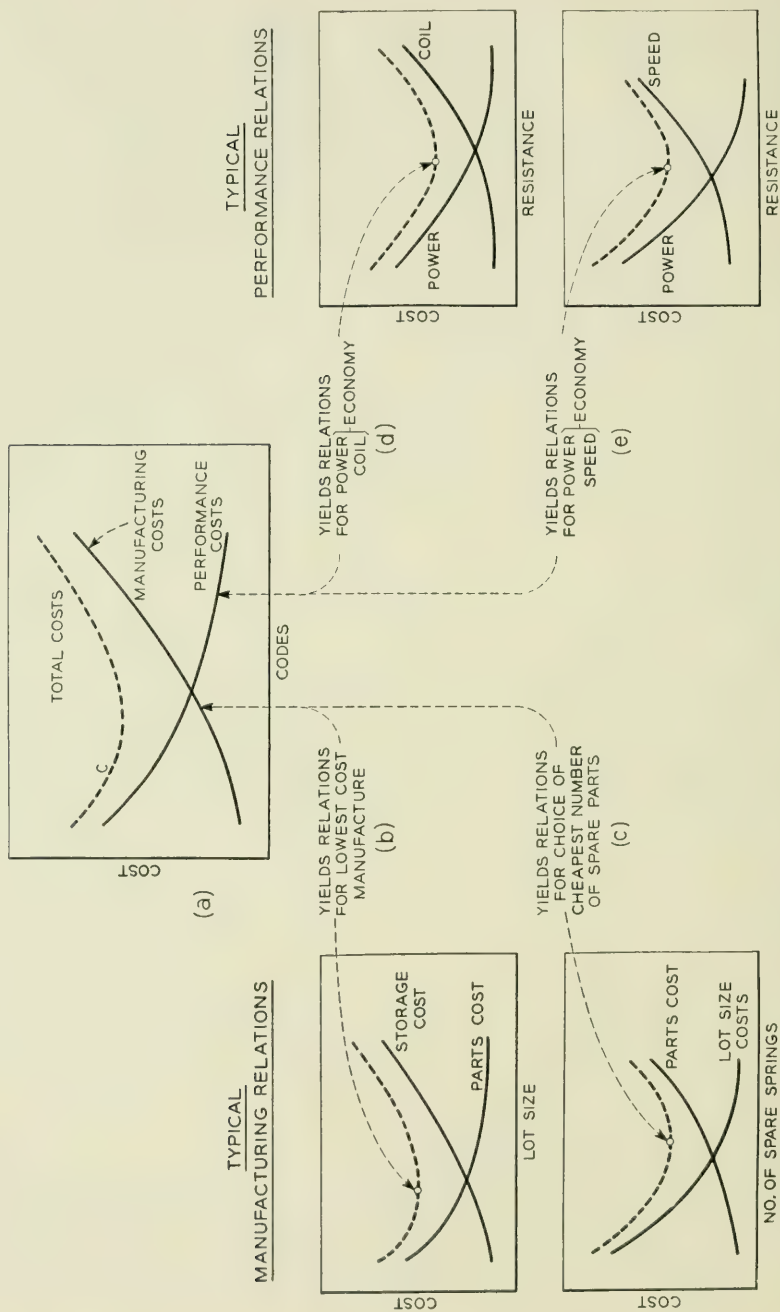


Fig. 1—Simplified view of procedure for gaining maximum system economy.

may apply (a) suitable windings to satisfy the various conditions imposed by a need for sensitivity, speed, lowest possible first cost, or some other of the many specialized requirements to be encountered; and (b) particular arrangements of contact springs which will provide for the desired functions in the circuit ranging from a single "make" up to complicated sequences such as 12 sets of "transfers." Every different combination of the basic parts is termed a "code," meaning another *kind* of relay built up from the parts common to the basic *type* in question. If an increasingly large number of such codes is encouraged, each tailored to some special circuit function, then the advantages of mass production begin to be lost as the total demand is split into more and more subdivisions, each with comparatively low demand. Each additional *kind* of relay is a problem in "coding economics," and can lead to excessive telephone office costs unless it is properly considered.

Satisfactory solutions to this coding problem in turn depend on a detailed knowledge of all the factors governing either the first cost of the relays, or the first cost of other features of the system, as controlled by the relays. When all such effects are properly stated, they can be compared and brought into economic balance. Though such studies soon branch out into many fields, the enormous dollar savings are strong incentives to do whatever work is needed.

Best economic balance in the system is formed through a series of optimizing steps, generally illustrated by Fig. 1. Graph (a) shows how the total cost of an office may vary with number of codes. The performance part of the cost diminishes with increasing codes because each code will more nearly satisfy the precise circuit needs. But the manufacturing costs of the relays will rise with increasing codes because the lot-sizes grow ever smaller and hence more costly. The lowest point on the summation curve represents a desirable goal. The two base curves of (a) in turn are built from detailed facts about parts costs, coil costs, power plant and equipment costs, and many other similar data. The curve of manufacturing cost, for instance, is built from graphs like (b) and (c) which tell how factory costs will vary. The curve of performance cost is built from graphs like (d) and (e) which tell how office costs will vary for things like power and speed. In each such case, important individual economies result by following similar optimizing steps. Procedures for the practical application of these ideas have been of particular help in recent redesigns of the No. 5 crossbar system to use the newly developed wire spring relay family.¹

¹ A. C. Keller, A New General Purpose Relay for Telephone Switching Systems, B. S. T. J., **31**, pp. 1023-1067, Nov., 1952.

In carrying out an actual study, a point of reference is needed first. This involves (a) fixing a standard against which all costs will be compared, and (b) expressing the value of all features in a common language. Once this has been done, every factor may be evaluated as an incremental "cost penalty," over and above the reference standard. Finally, in making comparisons, the reference standard will always subtract out, so that one needs only to add all incremental cost penalties to get the total cost of the design changes in mind.

The evaluation of all variables on a common basis is covered in Part I, which considers various manufacturing costs, power costs, and the cost of functioning time.

The remaining parts of this paper then develop relations for maximum economy in the switching system, for some important cases that arise in practice:

Coil design for maximum power economy.

Economical number of occasionally unused extra parts.

Economical adjustment for speed relay.

Coil design for maximum combined power and speed economy.

So far as possible, results are given in general form, permitting one to work from charts when considering specific application problems. Of further benefit are charts which permit the designer to decide the economic disadvantage of a design which may depart from optimum.

Design for systems economy through relay design includes many other important topics not considered here, as for example: how to make best use of molding, welding and other manufacturing processes; or how to design for long life, reliability, and other maintenance concerns. This article covers only some typical problems for which optimizing methods are readily applied.

PART I — EXPRESSING SYSTEMS PERFORMANCE AND MANUFACTURING VARIABLES IN A COMMON LANGUAGE

In the over-all switching system, one must evaluate various effects whose costs seem quite different, and a decision must be made as to what design course to follow, no matter how varied the conditions may be. Many of the most important such cases can be handled quite easily by a process of converting actual cost in the telephone plant back to an equivalent cost in terms of the factory production cost of each individual relay. Once all such costs are so stated, it is possible to examine the incremental effect of each change, and confidently draw conclusions. For example, if it can be stated that a change in a relay coil will save an amount of power that is worth 50 cents per relay equivalent first cost,

then there is no doubt about designing such a new coil if its manufacturing cost will increase by only 20 cents. So the relay designer needs to understand how various factors, from original manufacture to final application, may be stated on a cost basis that is comparable throughout. Among the most important special cases are the following:

1. Manufacturing cost of winding a coil,
2. Manufacturing cost as a function of annual demand,
3. Equivalent manufacturing cost of power consumed by relays,
4. Equivalent manufacturing cost of functioning time of a relay.

The method of evaluating each on the same basis will now be briefly outlined. In each case, it should be noted that since only comparisons between variable portions of the system are considered, only incremental values need be considered.

1.1 THE COST OF A WINDING

In comparisons between coils, certain common operations, such as soldering the leads, using and cementing spoolheads, etc., will always subtract out, leaving only the difference due to varying amount of copper, which is paid for by the pound, (or per ohm), and the number of turns, which depends on the speed of the winding machine and number of coils wound at one time. Thus, winding costs, which depend only on the electrical design, may be considered as varying only with the cost per ohm and cost per turn, as given in equation (1):

$$C_w = c_R R + c_N N, \quad (1)$$

where C_w = total cost of the variable part of the winding,

c_R = cost per ohm, which may be tabulated by wire size,

c_N = cost per turn of winding, also given in tables by wire size,

R = resistance of coil, ohms,

N = number of turns in winding.

Typical values of c_R and c_N for a particular kind of wire and coil are shown in Fig. 2.

1.2 COST AS A FUNCTION OF ANNUAL DEMAND

Various components of relays, as well as their assembly routines, occur in several variants of a general basic pattern. In the course of their manufacture it is necessary to stop the manufacture of one unit, reset the machine, and proceed with manufacture of another. Whenever this happens, production stops; and work must be done to reset the machine, all of which may be evaluated as a set-up cost (designated S). As more

and more variants are required in the over-all general structure, the number of set-ups will increase to the point where the economical manufacture of the part is seriously penalized. Thus a point will be reached where it is advantageous to make more parts than are immediately needed, and store the remainder, so as to avoid excessive cost of shut-downs. When the variable effects of (a) set-up costs and (b) inventory costs are fully considered, the ideal quantity to manufacture in one lot may then be found by comparing the results of (a) and (b), as will now be shown. The costs which correspond to these ideal quantities will immediately follow.

1.21 *Lot-Size Costs*

In manufacturing practice there are two outstanding expenses which may affect the cost of any particular item which is classed as a code, or kind, of the basic family of articles. These are the administration costs, and the set-up costs.

1.211 *Administration Costs*

For certain kinds of operations there is a considerable amount of paper-work, drafting, checking, etc. to maintain in normal up-to-date condition. Occasionally, this cost is enough to influence the cost of the

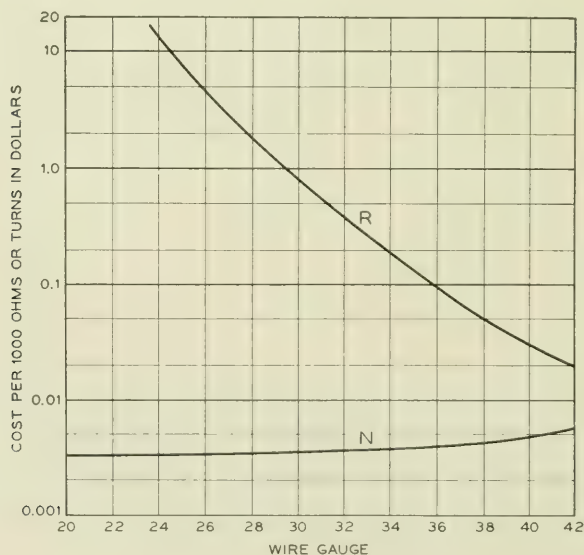


Fig. 2 — Typical winding costs.

item. The resulting individual effect may then be stated as

$$C_A = \frac{A}{n}, \quad (2)$$

where C_A = cost penalty per unit due to administration,

A = annual administration cost for maintaining one code in good standing,

n = annual demand for a given code.

1.212 *Set-up Costs*

If only one kind of part or assembly were needed, it could be continuously built in the same way, with no time lost for changeover to other parts, no particular bookkeeping necessary to control the proper flow of differing parts, and with more mechanized action. Usually this condition is far from realized in practice; nevertheless it represents the peak of manufacturing economy, and may be taken as a standard of reference for comparing all other less favorable conditions. The cost incurred, per item, due to its lot size, as against its cost if there were always but one lot, will be called the "lot-size cost penalty." The manufacturing lot-size cost penalty per part for any particular process, then, may be stated as follows:

$$C_L = \frac{S}{L}, \quad (3)$$

$$= \frac{S\ell}{n}, \quad (4)$$

where C_L = lot size cost per item,

S = cost of one "set-up",

L = size of lot for which one set-up is made,

$= n/\ell$,

n = annual demand for a given code,

ℓ = number of lots per year, or lot-frequency.

Equation (4) gives the cost over and above single-kind manufacture, so far as lot-size is concerned. Hence, when values for each of these variables can be established, the lot-size cost penalty for any particular process can be determined.

1.213 *Over-all Lot-Size Costs, Related to Annual Demand*

Combining the results of equations (2) and (4), the total cost penalty per part due to lot-size is

$$C_L = \frac{A}{n} + \frac{S\ell}{n}. \quad (5)$$

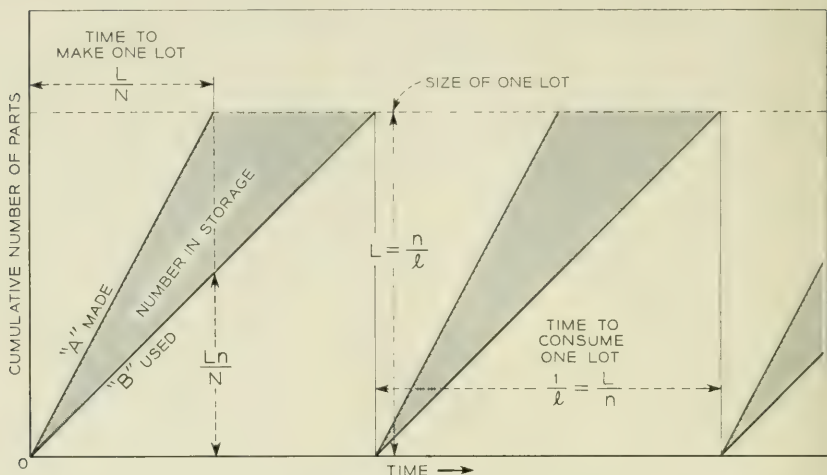


Fig. 3 — Lot-size cost penalty relationships.

1.22 Inventory Costs

When more parts are made than are immediately needed, it becomes necessary to store them until they are all used, when the process will be repeated. The resulting cost penalty, compared to no storage at all, may be stated analytically with the help of Fig. 3, which is a plot of production as a function of time. Line A represents the number of parts *made* in a lot as a function of time, while line B represents the parts *used* in the next manufacturing stage, as a function of time. The difference between the two lines at any time represents the number in storage, from which their value and hence the interest charges may be found.

The cost penalty per part due to storage is:

$$C = \frac{\text{Annual interest factor} \times \text{value of one part} \times \text{av. no. of parts in storage}}{\text{Annual production of this part}} \quad (6)$$

The following steps can be taken to supply values for this equation. From the figure, the maximum quantity of parts stored equals

$$L - \frac{Ln}{N},$$

or

$$\frac{n}{\ell} \left(1 - \frac{n}{N} \right).$$

The average number of parts stored is one-half this, or

$$\frac{n}{2\ell} \left(1 - \frac{n}{N} \right), \quad (7)$$

where N = rate of output of machine per year.

The penalty per part due to lot-size and administration was given in (5), and when added to its base cost C_1 gives the total value of one part as

$$C_1 + \frac{S\ell}{n} + \frac{A}{n}. \quad (8)$$

The penalty per part, due to storage, equation (6), is then the product of k/n times (7) times (8), where k = interest charges on stored parts expressed as a ratio. Then the storage cost penalty, C_s , is

$$C_s = \frac{k}{n} \left(C_1 + \frac{S\ell}{n} + \frac{A}{n} \right) \frac{n}{2\ell} \left(1 - \frac{n}{N} \right). \quad (9)$$

1.23 Total Cost Penalty

The entire penalty resulting from lot size charges and inventory charges may now be stated as a function of annual demand by summing equations (5) and (9). Writing q for $(1 - n/N)$, and rearranging terms, the total cost penalty in terms of annual demand is

$$C_P = \frac{kq}{2\ell} \left(C_1 + \frac{A}{n} \right) + \frac{kqS}{2n} + \frac{1}{n} (S\ell + A). \quad (10)$$

Equation (10) is the basic relation between costs and annual demand, and is of the general shape shown in graph (b) of Fig. 1.

It is of greatest interest to determine conditions where this curve has its lowest value, which occurs when $dC/d\ell = 0$. The result is

$$\ell_0 = \sqrt{\frac{kqn}{2S} \left(C_1 + \frac{A}{n} \right)}. \quad (11)$$

which gives the lot-frequency, ℓ_0 , for which total cost penalty is a minimum. When ℓ_0 is substituted for ℓ in equation (10), the resulting optimum cost, as a function of annual demand, C_0 , is given:

$$C_0 = \sqrt{\frac{2kqS}{n} \left(C_1 + \frac{A}{n} \right)} + \frac{kqS}{2n} + \frac{A}{n}. \quad (12)$$

Equations (11) and (12) furnish the means for deciding how to plan manufacture under differing conditions of annual demand, and how much the product is penalized even after planning is carried out as

efficiently as possible. Two illustrations will be given: for a comparatively low-cost part involving expensive set-up charges and no administrative charges, and for a more costly operation with comparatively inexpensive set-up costs but high administrative costs. For Case I assume a part for which the following constants apply:

$$\begin{array}{ll} k = 0.15 & C_1 = \$0.10 \\ N = 2,000,000 & A = 0 \\ S = \$10.00 \end{array}$$

In Cast II assume that

$$\begin{array}{ll} k = 0.15 & C_1 = \$1.00 \\ N = 2,000,000 & A = \$100 \\ S = \$2.00 \end{array}$$

The results for these two cases are given in Fig. 4.

Results such as these are of the utmost value both in manufacturing planning for lowest cost, and in applications planning to show when the lot-size penalties can be tolerated. However, no case is to be expected in practice where optimum conditions can be met more than a part of the time. What with unexpected rush orders, fluctuation of annual demands, rescheduling, possible moves, or breakdown of machinery, it is usually impossible to plan production to remain at all times at the optimum value. The variation from optimum cost that results from a departure from optimum lot size is thus of interest; it is readily stated

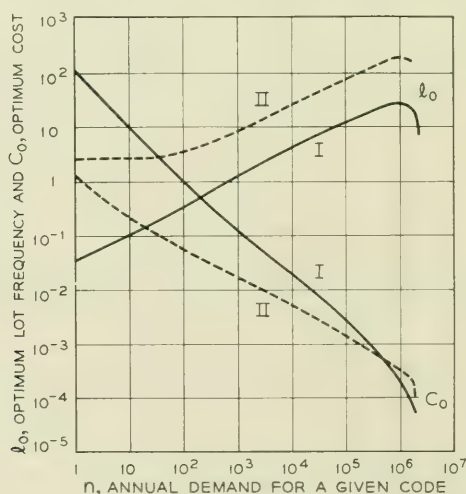


Fig. 4 — Optimum lot sizes and costs.

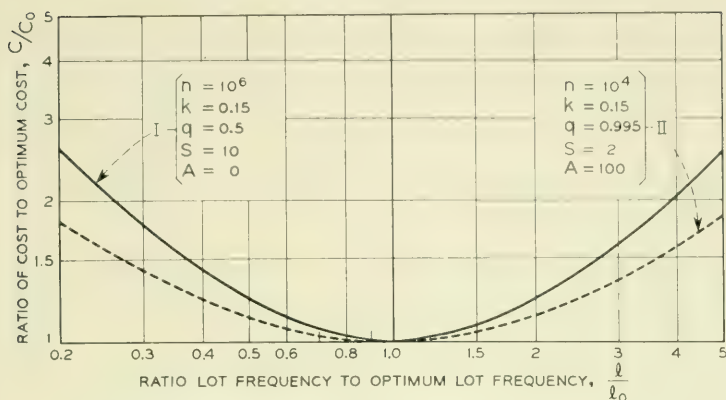


Fig. 5 — Cost variations with variations from optimum.

by rewriting equation (10) in terms of the values of ℓ_0 and C_0 just found in equations (11) and (12). The result is

$$C/C_0 = \frac{\ell_0/\ell + \ell/\ell_0 + D/\ell_0}{2 + D/\ell_0}, \quad (13)$$

where

$$D = \frac{kq}{2} + \frac{A}{S}.$$

This relationship is seen to depend on annual demands, and also administrative and set-up costs, which will be peculiar to each case. Therefore, it will be necessary to use the equation with appropriate values of these constants for each case that arises. Typical values for the two previous illustrations are shown in Fig. 5. Inspection of these curves shows that the cost penalty of departing from optimum lot-frequency is somewhat variable depending on the ideal lot-frequency for the particular case. However, the higher values of lot-frequency, which produce the highest penalties, are those most easily under control of the factory, and variations averaging in excess of 2 or 3 to one from ideal are not at all likely. When the ideal lot-frequency is low, manufacturing control is not so easy, but the cost penalty ratio of departing from the ideal is far less. From inspection of these curves, then, the value of C/C_0 must be judged by the conditions of each problem; a value of cost penalty amounting to about 1.25 C_0 is often found to be a quite satisfactory value to assume.

Thus, (a) a method is available to the manufacturing engineer to help plan optimum lot size, and (b) a method is available to the relay

and circuit designer for deciding how much his designs may be penalized by varying the quantities used. It consists of using the costs determined from equation (12), and modified by an amount indicated by (13), or in ordinary cases by arbitrarily increasing (12) by the factor 1.25.

1.3 THE EQUIVALENT MANUFACTURING COST OF POWER CONSUMED BY A RELAY

The largest part of the power plant used in telephone central offices is the 50-volt equipment needed for the switching apparatus. Because of its special voltage range and the requirements on its stability and reliability, it must be provided as a part of every central office to convert the normal power company voltages into the desired telephone values. This equipment involves generators, banks of storage batteries, and switchboard, bus-bar and control equipment; it is an expensive portion of every central office which, of course, it is desirable to minimize as far as possible. For every relay required in the central office, one must associate a small "chunk" of this power plant; so that each relay needed implies an associated investment in the power plant. In a later section the means for minimizing the power and relay costs are described, and since they will involve comparisons on the same basis, it is first necessary to state the value of the power plant in terms related to the amount of power consumed by any individual unit. The method of evaluating the power plant is given in this section.

The problem may be broken into two parts: (a) the equivalent first costs assignable to the power plant equipment, and (b) the equivalent first costs of the charges per kilowatt-hour paid to the power company. Then the figures are combined.

1.31 *The Power Plant*

Power plants furnished for central offices vary over a wide range of size and cost, depending on the size of the office and its estimated activity. This plant must of course be planned so as to carry the load during the period of peak activity, even though this occurs for only a small part of the time, so that in the long run the cost of the power plant is decided by the share of busy-hour power taken by each relay. To get such costs one must first find the cost of the plant as a function of total power requirements.

Present practice in power plant planning results in the purchase of basic power units which may be combined to cover certain ranges of power supplied over roughly a two-to-one range. Within this range, by

adding generators and batteries in parallel the needed capacity can be attained. Beyond the range, basically heavier machines are needed. The resulting installed price of such power plants has the character shown in Fig. 6. Since prices and installation charges change from year to year, the values shown here are given in qualitative form as illustrations—they must of course be evaluated as concisely as possible when considering new central office designs. As seen in the figure, the power plant price varies in two ways: in fairly big jumps as basic plant size is changed, and in smaller but rather uniform manner as the basic size varies within its lower and upper limits.

From Fig. 6 it is seen that for each basic size of plant the rate of increase in price with small increases in capacity is about the same for all plants. The average rate for this incremental power capacity, corresponding to the expense if relatively small power changes are made in any one basic type of power plant will be called a . As the power requirements of the office change beyond a certain point, it is necessary to install basically larger or smaller equipment with correspondingly different base prices. At the points of maximum capacity, there is an abrupt change in the basic price of the plant, amounting often to many thousands of dollars. These abrupt increments, designated A , represent a lump sum of money that can be saved whenever the power plant size can be re-



Fig. 6 - Installed value of power plants.

duced to the next lower-range unit. Such a saving can be realized only part of the time, but it is desirable to attempt to weight its value so that over the telephone system as a whole there is adequate encouragement to strive for power savings.

The method of weighting is complicated by such factors as the fraction of power that may be saved in an office, by the likelihood of certain size offices being representative, and by the extent to which offices are designed with safety factors for growth and overload. Nevertheless a reasonable estimate can be made by assuming differing amounts of the office power can be saved, and by assuming that all sizes of office are equally likely. This gives the result that the over-all incremental value of power is

$$k = \frac{C}{W} = a + \frac{2AW_0}{(1-p)(W_2^2 - W_0^2)}, \quad (\text{dollars per kw}), \quad (14)$$

the average value of power plant per kilowatt saved over the range between W_0 and W_2 , taking into account the proportionate number of plants that can be converted to the next cheaper size, where

C = total price saving,

W = total power saving,

p = fraction of power saved, not to exceed about .5, (above which point two steps of base price saving might be realized),

W_0 = lower limit of range,

W_2 = upper limit of range,

a = incremental price of one size power plant, assumed to be about constant for all plants,

A = difference in base price to next lower plant, as may be tabulated for any particular case.

By these methods, incremental values of installed power plant may be found; it is even possible in many cases to assume an average slope for the family of curves in Fig. 6, without greatly changing the results in practice.

It is now desired to convert these figures of installed expense to figures comparable to manufacturing cost, in the factory, of the relays which will consume this power. This may be done through standard economic-study practice by which first cost is converted to annual charges through a knowledge of charges for interest, taxes, depreciation, etc. As these factors are not necessarily alike for power equipment and relays, the factors used will differ to suit the problem. For the present study, with recognition of the accounting practices then in effect, it was found that

on a basis comparable to relay manufacturing cost, savable power was worth between \$500 and \$2,500 per kilowatt, depending on size, savable power, etc.

Now these figures must be associated with the power needed by any particular relay. The total switching power during the busy part of the day, which is the period determining the size of the power plant, is the sum of all the individual relay power drains at any particular moment. This in turn may be taken as the sum of the power required by each relay times the probability that it will be energized at any particular time, which is taken as the fraction of the busy hour that the relay is expected to be energized. For every relay, this ratio can be determined with some certainty, by consultation with the circuit designers.

Summarizing, the switching power capacity required by the office may be stated as the sum, for all relays, of power w for each relay, times m , the fraction of the busy hour that it is energized (i.e., hrs. per busy hr.). The cost of the power plant capacity required for each relay is then

$$C = kmw \text{ dollars,} \quad (15)$$

where k is the equivalent value of power.

Very often it is desirable to state this cost on the basis of the annual power drain. In this way, the plant investment can be easily correlated with the costs paid to the power companies. In telephone offices, the annual power has been found to correspond to 3,000 times the power consumed in a busy hour, and as a matter of convenience power drain in the central office is often stated in terms of the energy per year based on a 3,000-busy-hour year. Then the ratio of use during the busy hour may also be stated as

$$m = \frac{h}{3000},$$

where h = hours per year energized.

Writing $E^2/1000R$ for the power in kilowatts drawn by any relay, the equivalent first cost of the power plant assignable to the relay is then

$$\begin{aligned} C_{\text{power plant}} &= \frac{E^2}{1000R} \times \frac{kh}{3000}, \\ &= c_{P_1} \frac{E^2 h}{R}. \end{aligned} \quad (16)$$

For this case, in other words, the equivalent first cost of power plant has a value c_{P_1} dollars for each watt-hour-per-year of switching power required.

1.32 *Cost of Power Purchased from Power Companies*

The power used by the switching apparatus corresponds to a larger amount of power furnished by the power company, since the power machines are not 100 per cent efficient. Its cost may be stated as

$$c_P = \frac{c_0}{e},$$

where c_0 = average marginal cost of power furnished,
 e = conversion efficiency.

The annual power bill will then be

$$\frac{c_0}{e} \frac{E^2 h}{1000R} \text{ dollars.}$$

Converting from annual charges to first cost, through the same steps as above, gives

$$C_{\text{power}} = c_{P_2} \frac{E^2 h}{R}, \quad (17)$$

or c_{P_2} dollars per watt-hour-per-year.

1.33 *Total Power Cost*

With both power plant costs and purchased power costs now stated on the same basis, they may be combined to give the total power cost, C_P :

$$C_P = (c_{P_1} + c_{P_2}) \frac{E^2 h}{1000R} = \frac{c_P E^2 h}{R}, \quad (18)$$

where c_P represents the dollars per watt-hour-per-year for power consumed by a particular relay, and is the equivalent first-cost value of power.

Thus a method is available to find the value of the power that magnet coils may need — to be balanced against the first cost of the coils (Section 1.1). Procedures for best results are given later.

1.4 THE EQUIVALENT MANUFACTURING COST OF THE FUNCTIONING TIME OF A RELAY

Another large part of the investment in every crossbar-type central office is the common-control equipment, most important of which is the "marker." This equipment has the sole duty of selecting the proper path for each call, and sending it on its way. After each such function,

the equipment restores to normal, ready to serve the next call. It takes a marker about a half-second to do its job; nevertheless many markers may be needed to handle the traffic during each day's heaviest calling period. Since the cost of each marker is measured in the tens of thousands of dollars, it becomes a matter of great importance to shorten the marker work time — in other words, to shorten the time of each relay in the chain of marker events. Since faster relays are usually more expensive it is important to find the dollar value of speed so as to guide an intelligent design of each speed relay.

As in the case of power, it has been found possible to state the value of speed in terms equivalent to the manufacturing costs of relays. This may be done through a knowledge of the number of markers and associated equipment needed in relation to their work time, and their cost. Such relationships are shown in Fig. 7, which gives a typical curve for the markers per line needed to handle the traffic. Then, when the value of the marker is known, the value per line per millisecond, c_l , follows. Corresponding to this, a value per millisecond per marker, c_M , can then be found, which varies inversely as the number of markers needed:

$$c_M = \frac{c_l}{p},$$

where p = number of markers per line that are needed. This figure is the value of a millisecond for any complete event or series of events

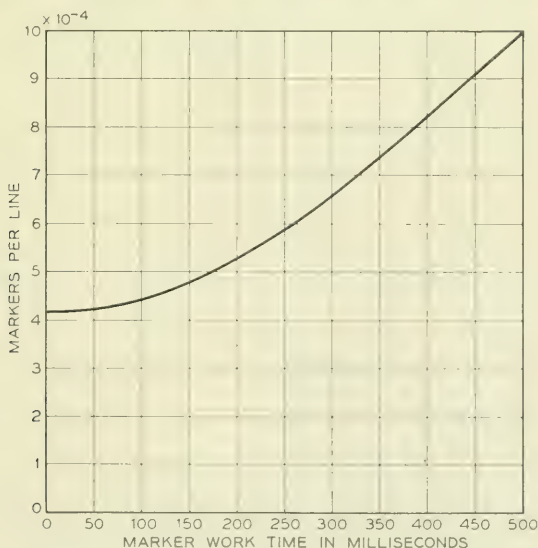


Fig. 7 — Markers as a function of work time.

within the marker which directly affects its working time, referred to as a "major event."

1.41 *The Value of Speed per Event*

Within each major event there may be a whole cycle of operations happening partly in parallel and partly in sequence. Each separate action is called an "event." In the case of events happening in parallel, the circuit designer will usually know whenever one path is in outstanding control of the work time as a whole. Whenever such is the case, it is a "major event," and the other paths may be ignored. Often, however, the times for each of several paths will be about the same, and special consideration is needed to determine their value. In breaking down these operations so as to arrive at the cost per relay, one must evaluate the events within any major event as in the region marked in Fig. 8. Supposing that the number of subevents within a major event are labelled a, b, c , etc. for each of the several parallel paths, then the total number of subevents is $a + b + c + d + \dots$. The total time for the major event is t , which has the dollar value, c_t/p , per millisecond, as just previously given.

Along any path, A , of events, the total time for the events is the time taken by the group as a whole. If this total time is t , then the effective time per event along any path is

$$\begin{aligned} t_A &= t/a \text{ for path } A \text{ (seconds),} \\ t_B &= t/b \text{ for path } B \text{ (seconds),} \\ &= \text{etc.} \end{aligned}$$

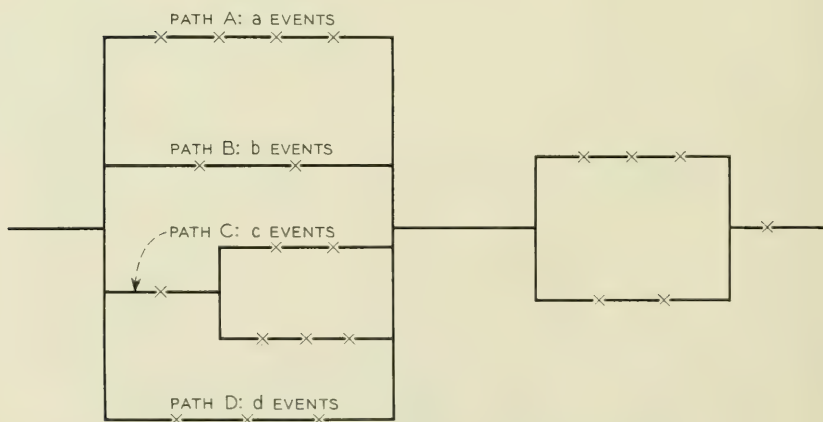


Fig. 8 — Typical chain of events in controller operation.

The worth of the time taken for the group as a whole is then

$$C_M = c_t/p = c_t a t_A/p = c_t b t_B/p = \text{etc.}, \text{ dollars.}$$

The value per event is usually of principal interest. Since there are $(a + b + c + \dots)$ events in any one "major event," then the total value per event C_E is

$$\begin{aligned} C_E &= \frac{C_M}{(a + b + c + \dots)} \text{ dollars per event,} \\ &= \frac{a}{(a + b + c + \dots)p} c_t a, \quad \text{for path A,} \\ &= \frac{b}{(a + b + c + \dots)p} c_t b, \quad \text{for path B,} \\ &= \text{etc.} \end{aligned}$$

Then the value of an event, per millisecond, is

$$c_E = q_A c_t/p = q_B c_t/p = \text{etc.}, \text{ dollars per event per millisecond.}$$

$$\text{where } q_A = \frac{a}{a + b + c + \dots}, \quad q_B = \frac{b}{a + b + c + \dots}, \quad \text{etc.}$$

In general, then, the value of an event is

$$C_E = c_E t_E \text{ dollars,}$$

where t_E = the duration of the event,

$$c_E = \left(\frac{q}{p} \right) c_t \text{ dollars per millisecond.} \quad (19)$$

The quantity q will be called the "path factor." It is seen to simplify to a value of unity when a major event is involved, ($q = a/a$), and in all other cases to have a value less than this.

1.42 The Value of Speed, per Relay Operation

Each event may consist of the action of one relay out of several possible alternates. This means that several relays must be purchased for each such event, though only one determines the time. Calling w the number of relays provided per marker to perform one function, each function being designated by subscript corresponding to path and se-

quence, then the value per relay of one millisecond is

$$\begin{aligned}
 c_r &= \frac{q_A c_t}{w_{A_1} p}, & \frac{q_A c_t}{w_{A_2} p} & \text{ etc., for path } A, \\
 &= \frac{q_B c_t}{w_{B_1} p}, & \frac{q_B c_t}{w_{B_2} p} & \text{ etc., for path } B, \\
 &= \frac{q c_t}{w p}, & & \text{ in general, (dollars per millisecond).}
 \end{aligned} \tag{20}$$

These equations give the general expression for the value of functioning time in terms of cost. They permit the direct comparison with cost of a relay design, including its coding cost, manufacturing cost, or power drain cost, in deciding on how much investment is warranted in switching apparatus in order to shorten the marker work time.

1.43 *Simple Examples of Relay Design for Speed*

The above relations make it possible to rapidly evaluate design changes for speeding up relays, or to state the value of seeking a design change that will make speed possible. Several examples will be given.

(a) *Relay Involved in a Major Event*

Suppose the following typical conditions are given:

$$\text{No. of markers per line} = p = 0.0006.$$

$$\text{Path factor} = q = 1.$$

$$\text{Relays per marker function} = w = 1.$$

$$\text{Value per millisecond per line} = c_t = \$0.04.$$

Then value per millisecond per relay is

$$c_r = \frac{0.04 \times 1}{0.0006 \times 1} = \$67.$$

In this case a very large investment per relay could be justified in order to gain small savings in time.

(b) *Relay Involved in a Subevent*

Suppose the following typical conditions are given:

$$\text{No. of markers per line} = p = 0.0006.$$

$$\text{Path factor} = q = \frac{2}{8} = 0.25.$$

Relays per marker function = $w = 2$,

$$c_t = 0.04.$$

Then value per millisecond per relay is

$$c_r = \frac{0.04 \times 0.25}{0.0006 \times 2} = \$8.33.$$

In this case, too, it is worth a considerable manufacturing cost penalty to shorten the functioning time of the relay.

(c) *Relay Involved in a Subevent—Release*

A particularly common form of subevent is the release of numerous relays at the end of one phase of marker action. In this case, assume values of the constants are:

$$\text{No. of markers per line} = p = 0.0006.$$

$$\text{Path factor} = q = 0.02.$$

$$\text{Relays per marker function} = w = 5.$$

$$c_t = \$0.04.$$

Then value per millisecond per relay is

$$c_r = \frac{0.04 \times 0.02}{0.0006 \times 5} = \$0.266$$

Thus even in this case, an appreciable first cost value is associated with the releasing speed of each relay.

1.5 REVIEW OF EQUIVALENT FIRST COST EXPRESSIONS

The previous sections have shown how some of the variable portions of the relay design that affect its performance may be stated in terms of first cost. Likewise methods are given for evaluating sensitivity and speed performance in the same terms. These may be summarized as

Cost of a coil:

$$C_w = c_b R + c_s N. \quad (1)$$

Cost of an item as a function of demand:

$$C_o = \sqrt{\frac{2kqS(C_1 + A/n)}{n}} + \frac{kqS}{2n} + \frac{A}{n}. \quad (12)$$

Cost of power:

$$C_p = \frac{c_p E^2 h}{R} \quad (18)$$

Cost of functioning time:

$$C_r = \frac{qc_t}{wp} \quad (20)$$

Examination of these equations will show that, in nearly every case, a cheapening of the cost of the structure leads to more expensive (i.e., poorer) performance, and vice versa. However, the availability of these relations, together with accurate values for the constants, gives the information needed to vary the relay design to the point where maximum over-all economy (i.e., the cheapest central office) can be realized.

In the following sections, some examples of methods for finding this optimum point, and how much it is worth to find it, will be given.

PART II — DESIGN FOR MAXIMUM POWER-PLUS-COIL ECONOMY

Among the more common problems facing the designer of switching circuits is the selection of the relay magnet which gives best economy in use. On the one hand, equation (1) has shown that the relay coil cost will be less if the resistance is less (but power consumed is more). On the other hand, equation (18) shows that the equivalent first cost of that part of the power plant furnished to operate the relay increases as the resistance is reduced (but coil is more expensive). These opposite effects are indicated in Fig. 9 where the curve labelled C_w shows how wind-

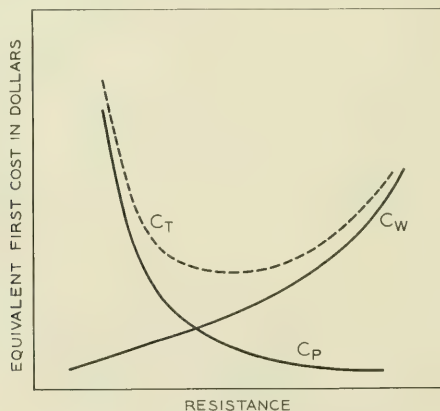


Fig. 9 — Power and winding costs.

ing cost for a certain wire gauge varies while the curve labelled C_P shows how power plant cost is changing. The desirable end result is a coil with resistance value to permit the sum of these to be as low as possible, as in the curve marked C_T . The following methods will give such a design.

The value of C_T , made up of power cost, C_P , and winding cost, C_W , represents the cost of all the parts of the central office which are influenced by a change in resistance of the coil. Thus we wish to minimize this equation:

$$C_T = C_P + C_W,$$

where, from the previous relations,

$$C_P = \frac{K_1}{R},$$

$$C_W = K_2 R,$$

$$K_1 = c_P h K^2,$$

$$K_2 = c_R \left(1 + \frac{c_N}{c_R} \frac{N}{R} \right).$$

The resulting expression for this total cost is

$$C_T = \frac{K_1}{R} + K_2 R. \quad (21)$$

In this equation K_1 is a constant, but K_2 varies both with resistance and wire size, as the latter affects the values of c_N , c_R , and the quantity N/R , which is determined by the coil design equation

$$R = AN(h + d),$$

in which, using the special notation common to coil design problems:

N = turns in winding = $h\ell/K$,

R = coil circuit resistance,

A = π times the resistivity of wire in ohms per inch length, given in tables for each size of wire,

h = depth of coil space occupied by wire,

ℓ = length of winding space,

d = inside diameter of coil,

K = effective cross-sectional area required by one turn of wire, given in tables for each size.

For a given relay, the dimensions ℓ and d are fixed, and only the wire size and the coil depth h can be varied. For a given wire size, therefore, the values of h and N are fixed by the resistance. As C_N and C_R are known quantities, the values of K_2 can be determined and plotted against R for each wire size, as illustrated in Fig. 10. Such a chart applies to a particular relay, as the curves depend on the fixed values of ℓ and d applying. Here it can be seen that for a given resistance the finest gauge wire is the cheapest. Because of this, and since power costs are minimized with the highest resistance, the finest gauge wire would always give the cheapest coil. However, this becomes a theoretical case. As finer and finer wires are used, $A = \pi\sigma$ becomes larger, and thus N/R or NI/E becomes smaller. In all relay design, an ampere-turns requirement must be met. Thus a practical solution to this problem of optimum costs is to choose the finest gauge wire which will meet the ampere-turns requirement.

If the same relay is assumed as for Fig. 10, the variations of ampere turns with resistance are as shown in Fig. 11. Starting with the NI required, the wire size can be determined from a plot like Figure 11. Then K_2 is determined from a plot like Fig. 10.

Now, with the gauge of wire chosen, K_2 is almost a constant and can be considered independent of R . Thus, upon differentiating (21), C_T will be found to be a minimum when R has a value designated by

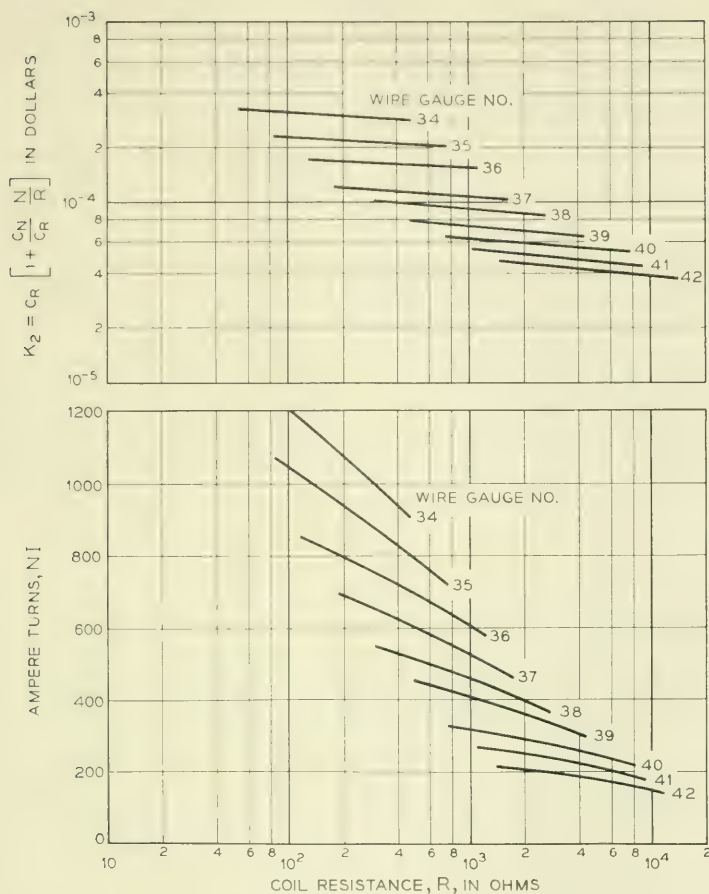
$$R_0 = \sqrt{K_1/K_2}. \quad (22)$$

The most economical coil, resulting from use of a resistance R_0 , will have a system cost of

$$C_0 = 2\sqrt{K_1K_2}. \quad (23)$$

As can be seen by the slopes of the curves on Fig. 11, if the optimum resistance, R_0 , is larger than the resistance value at the desired ampere turns, the NI requirement will not be met. Then either the selection of R_0 must be made for the next coarser wire, or the resistance value chosen at just the required NI , whichever is cheaper. Conversely, if there is a very large NI margin, a finer gauge wire could be tried. In all, two trials should give the optimum values.

A procedure has now been outlined for choosing the coil design for optimum power-plus-winding cost. Curves similar to Figs. 10 and 11 should be drawn for the applicable relay parameters. Then the K_2 value is obtained for the finest gauge meeting the ampere-turns requirement, and R_0 and C_0 found. Then, the NI value applying for R_0 must be checked and adjustments made as above, if necessary.



Figs. 10 and 11 — K_2 and NI values versus resistance for various wire gauges for a typical relay.

COST WHEN OPTIMUM RESISTANCE CONDITION IS NOT SATISFIED

There will, of course, be cases in practice where the optimum resistance is not used, for such reasons as standardizing of certain coil sizes, need for speed, or insufficient winding space. The penalty in cost from departing from the ideal winding size may be found by comparing the cost for any value of resistance, (equation 21), with the cost if resistance were optimum, (equation 23):

$$\begin{aligned} \frac{C}{C_0} &= \frac{K_1/R + K_2R}{2\sqrt{K_1K_2}}, \\ &= \frac{1}{2} \left(\frac{R_0}{R} + \frac{R}{R_0} \right). \end{aligned} \quad (24)$$

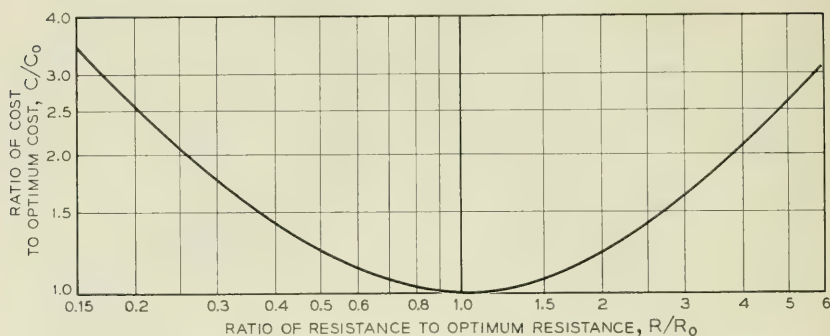


Fig. 12 — Cost penalty when resistance is not optimum.

Thus, the fractional change in cost caused by any departure from the best resistance is readily estimated, expressed in ratio form as R/R_0 . This relation is shown in Fig. 12.

SUMMARY

In summary of Part II, methods have been given for deciding on the magnet coil design which will provide the cheapest net cost in the switching system. When the conditions of use are known, a series of charts may be prepared such as the one shown, from which the applications engineer may decide the proper coil design, and how much it may be worth to the system. As experience will show, the cost of departing from an optimum design may run well over \$1.00 per individual relay, with the result that aggregate savings in the whole central office may run to thousands of dollars compared to the cost when this problem is ignored.

New relay development may be guided by the use of these same relations, since through a process of considering various hypothetical designs of various sizes, work outputs and magnetic qualities, the potential optimum costs may be found. Design of the recently adopted wire spring relay was very materially guided by considerations of this sort.

PART III—CHOICE OF CONTACT SETS FOR MAXIMUM ECONOMY

One of the many problems involving lot-size costs of relays is the number of standardized sets of contacts that should be provided: i.e., how many kinds of contact arrangements give maximum economy for the product as a whole? By providing only a few choices of kinds of contacts, for example, the annual output is less divided and there are fewer setup charges, with the result that the initial manufacturing economy is large. However, because of the small selection of kinds of contact sets, there will result many cases where extra contacts are provided at greater than

needed cost. The problem is how best to combine these opposing factors.

In illustration, consider a relay which is to be equipped with molded spring arrangements, capable of providing any number of "transfers" from 1 to 12, each of which is used in approximately equal numbers in the system. At the one extreme, only one molding might be provided, always equipped with 12 transfers. At the other extreme 12 different moldings might be provided, covering each individually needed quantity of "transfers." In the case of the single molding there is no lot-size penalty at all, but on the average a large penalty in surplus contacts; while for the twelve kinds of moldings, lot-size penalties corresponding to one-twelfth of the possible annual output are incurred, but no spare parts at all will be needed. Such a problem may be treated as given below.

Let the number of kinds of spring sets chosen be designated by v . Then the annual output of each kind is the total output N divided by v , or

$$n = \frac{N}{v}. \quad (25)$$

Now it was shown in Section 1.2 that the cost penalty for making things in lots is given by equation (12). For the present problem, this may be expressed in terms of v , by substituting equation (25) into equation (12). The resulting lot-size cost penalty is then

$$C_0 = \frac{v}{N} \left[\sqrt{\frac{2kSN}{v} \left(1 - \frac{1}{v}\right) \left(C_1 + \frac{Av}{N}\right)} + \frac{kS \left(1 - \frac{1}{v}\right)}{2} + A \right]. \quad (26)$$

The penalty due to providing unneeded springs is the dollar value of the average number of spare springs. The average number of spares is approximately

$$E = \frac{12 - v}{2v}.$$

Thus for only one kind furnished, always 12 transfers, the number of spares can vary between 0 and 11, an average of 5.5, as given by the equation. If there were eight kinds, consisting possibly of moldings of 1, 3, 5, 6, 7, 8, 10 and 12 sets of transfers, the average number of extras would be 0.25. The dollar penalty compared to no extras ever needed is

$$C_E = C_1 \left(\frac{12 - v}{2v} \right). \quad (27)$$

and the sum of this equation and that for lot size represents the system cost for any one condition.

Fig. 13 shows this variation for an assumed case for which

$$N = 500,000 \text{ parts per year.}$$

$$k = 0.15.$$

$$S = \$100.$$

$$C_1 = \begin{cases} 0.20 & \text{for spring set cost.} \\ 0.005 & \text{for spare spring cost.} \end{cases}$$

$$A = 0.$$

Here it is found that the greatest benefit comes from using in the neighborhood of six kinds of spring sets. It is further clearly seen that the system as a whole will not be particularly penalized if the number of sets chosen is a few different from this quantity — but a penalty of almost two cents per relay, or about two thousand dollars per year would be incurred by erring too far toward the extreme.

PART IV — CHOICE OF SPECIAL ADJUSTMENT FOR MAXIMUM SPEED ECONOMY

Speed can be an exceedingly valuable property of a relay, as shown in Section 1.4. There are cases for example where every saving of 1

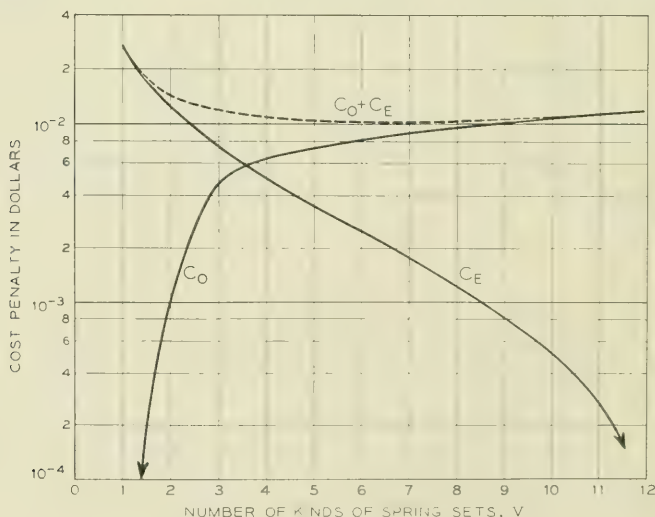


Fig. 13 — Optimum number of spring sets.

millisecond in working time will reduce the average quantity of equipment needed by an amount equivalent to \$35, or even more. For such cases a considerable sum of money can be spent on each relay in order to make it faster. The following approach helps guide the circuit and relay designer in the choice of relay they should make in order to gain the greatest over-all advantage.

It is well known to relay designers that one of the most important design parameters controlling operating time of a relay is the spacing between the contacts when they are at rest. This space cannot be less than a certain small distance of about 0.005" to avoid contact failures due to vibration, buildups, sparkover, etc. However, because it is costly to hold adjustments accurately to a few thousandths of an inch, the maximum spacing is often quite large, sometimes as much as 0.050". The spread in this distance is primarily a matter of economics; if it were clear that more money could be spent on the relay to narrow down this spacing while gaining material reductions in operating time as a result, then some closer adjustment would be specified. Such a problem is readily treated by summing (a) the cost penalties for differing values of this spacing and (b) the cost penalties of the corresponding functioning time, and finding the optimum condition that results. One such case is shown below.

The cost penalty due to changing spacing may be estimated by considering various actual values, and determining what procedures would be used in the factory in each case. Such a cost study can be made in the factory. In illustration, assume a particular relay type deemed to suffer no cost penalty if its armature stroke is allowed to take on any value up to 0.050", but which cannot be less than 0.010" without impairing performance. The actual cost per unit of various kinds of manufacturing procedures might be found according to the hypothetical conditions given in Table I. Thus a curve can be plotted of the cost penalties for each setting of gap.

The design may also be studied for the influence of the gap settings on speed. The operating time of the magnet as a function of spacing is readily determined by experiment, and may also be checked against known operating principles, to be sure the figures are reasonable. For a typical case, operating times corresponding to various spacings might be as given in Table II. At the same time, the value of this functioning time is given by equation (20) to be

$$C_r = \frac{qc_t}{wp}.$$

TABLE I — COST PER UNIT OF VARIOUS KINDS OF MANUFACTURING PROCEDURES

Assumed Max. Stroke	Manufacturing Procedure	Cost
0.050"	Acceptance of wide dimensional tolerances on all parts, and no adjustment permitted	\$0
0.045"	As above, but crude adjustments added	0.05
0.040"	Closer dimensional tolerances, no adjustment	0.07
0.034"	As just above, with rough adjustment	0.10
0.025"	Very close tolerances on all parts, no adjustment	0.15
0.020"	As just above, with touch-up permitted	0.20
0.015"	Screw adjustment added, (more expensive parts, but simpler adjustment)	0.35
0.010"	Precision setting of all parts	0.65

Assuming for this hypothetical case that the values applying are

$$c_t = \$0.05,$$

$$w = 100,$$

$$q = 0.2, \text{ and}$$

$$p = 0.0006,$$

then the cost of the resulting time is found in the last column of Table II.

A summation of these two sets of cost penalties is given in Fig. 14, which clearly shows that the ideal value of stroke is about 0.020", and that that it may be permitted to vary between the limits 0.015" and 0.025" without a serious economic penalty.

In summary, a method has been indicated for deciding how important it may be to build in, or omit, expensive design features which have an effect on functioning time. An illustration has been given for an assumed common control circuit and an assumed change in a design feature of the relay that importantly affects the time. The method, however, is equally

TABLE II — OPERATING TIMES CORRESPONDING TO VARIOUS SPACINGS

Assumed Max. Stroke	Functioning Time (Milliseconds)	Equivalent First Cost Value of Time
0.050	6.1	\$1.02
0.045	5.6	0.933
0.040	5.2	0.867
0.035	4.75	0.792
0.025	3.75	0.625
0.020	3.2	0.533
0.015	2.55	0.425
0.010	1.9	0.317

applicable to other speed problems, once the value of speed is known, and once the influence of the design change and its cost are known.

PART V — CHOICE OF WINDING FOR MAXIMUM COMBINED ECONOMY OF SPEED AND POWER

In the selection of magnet designs for use in circuits where speed is important, much can be gained by modifying the mass, the stroke, the force against the stop, etc., but when all these devices have been exhausted there remains the possibility of supplying more power to the magnet coil. If this is done in enough cases, an over-all penalty in central office cost is incurred through the added over-all power plant capacity that may be needed. As already seen in Section 1.3, this cost penalty is proportional to the base value of one watt-hour-per-year, the voltage squared, and the hours per year energized, but is inversely proportional to the resistance. In telephone usage, the voltage is usually fixed at 50 volts, and equation (18) may be rewritten as

$$C_P = \frac{k_P h}{R}, \quad (28)$$

where

$$k_P = c_P L^2.$$

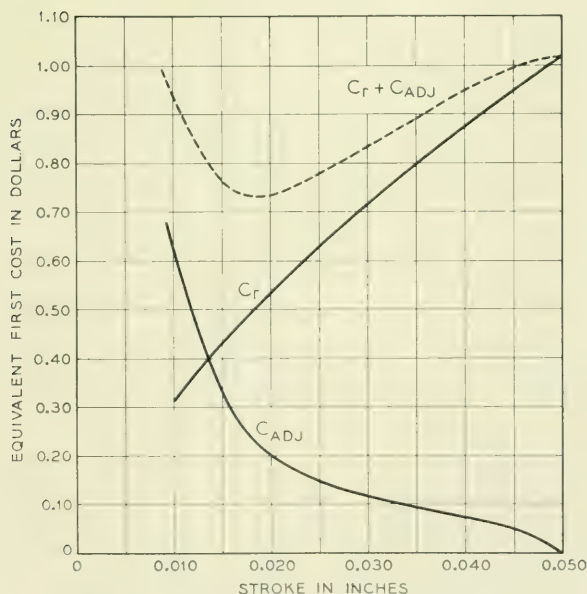


Fig. 14 — Optimizing a speed relay design.

The above expression gives the equivalent first cost of the power taken by one relay. For each relay operation, or event, however, there may be several relays energized. Then the total cost penalty per event, due to power, is

$$C_P = \frac{k_P h g}{R} \text{ dollars,} \quad (29)$$

where g = number of relays energized, per marker, for the particular function in question even though only one appears in the time sequence. For the power cost to be compared with the cost due to speed, this cost must be spread over the number of relays, w , involved in the function. Then

$$C_P = \frac{k_P h g}{w R} \quad (30)$$

This is the equivalent first cost penalty incurred against each relay, due to its power consumption. As R increases, it is seen to decrease.

However, as R increases, the operate time increases. For speed relays, the time t has been found to be a complicated function of the resistance, varying approximately as $R^{1/3}$. Thus when R increases, time increases, and the equivalent first cost of this action time increases, as already seen in Section 1.4.

As power is changed, then, there results two oppositely varying costs — one due to power, and one due to speed. The problem is graphically shown in Fig. 15, where Curve A shows the typical cost variation due to power drain and Curve B shows the variation in cost due to speed. The point where the sum of the two is a minimum is the ideal point to operate, assuming that the coil costs are about equal in each case. Curve C shows how the total cost will vary. Actually, it may be necessary to add the effects of a third variable: the cost of the coil itself as it is redesigned for different power consumption. This can be easily done by the methods shown above, but is ignored in the present treatment for the sake of simplicity. In many actual cases, coil cost has been found to be a much smaller factor than the other variables.

CHOICE OF OPTIMUM COIL

In practice the above result may be obtained by a knowledge of how a specific magnet design will vary in operate time as its coil resistance is changed. The values may be substituted in the following equation for total cost:

$$\begin{aligned}
 C &= \frac{q c_l t}{w p} + \frac{k_p h g}{w R}, \\
 &= \frac{g}{w} \left(\frac{q c_l t}{p g} + \frac{k_p h}{R} \right).
 \end{aligned}
 \tag{31}$$

The expression may be evaluated when a relation between t and circuit resistance R is established, usually a matter for experiment in any given case.

Illustration

For a particular case let us assume that $p = 0.0006$, so that equation (31) is

$$C = \frac{g}{w} \left(1670 \frac{g}{g} c_l t + \frac{k_p h}{R} \right). \tag{32}$$

Typical values of operate time, t , for a speed relay in 50-volt telephone applications are given in Table III. From this information, one may readily evaluate equation (32) to cover the practical range of conditions that are met in service. This has been done, in Fig. 16, to give the total

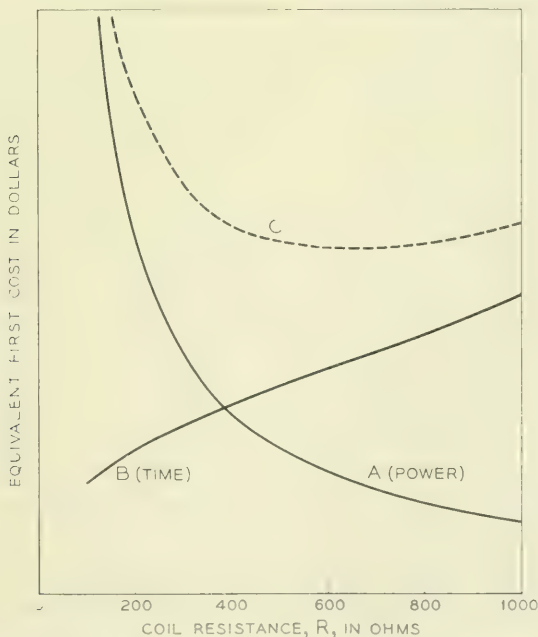


Fig. 15 Optimizing costs of speed plus power by relay coil design.

TABLE III — TYPICAL VALUES OF OPERATE TIME

Coil Resistance (Ohms)	Optimum Time (Milliseconds)
100	2.45
200	3.1
400	4.15
600	5.0
800	5.8
1000	6.7

cost penalty for the case of very short holding time ($h = 100$ hr. per year) and for long holding time ($h = 2000$ hr. per year). The optimum coil resistances for these cases are seen to lie between 200 and 400 ohms for the long holding time condition, and to be as low as possible for the case of short holding time. As a result we see an appreciable economic incentive to plan relay designs which are capable of operating on low resistances. Further study of heat dissipation, operation with holding windings, contacts to withstand heavy currents, and other devices to gain fast action, therefore take on added importance to the relay designer.

PART VI — CHOICE OF THE ACTUAL CODE

With the background of the previous sections, the systems designer can now decide what codes he needs to use in his circuits so as to be

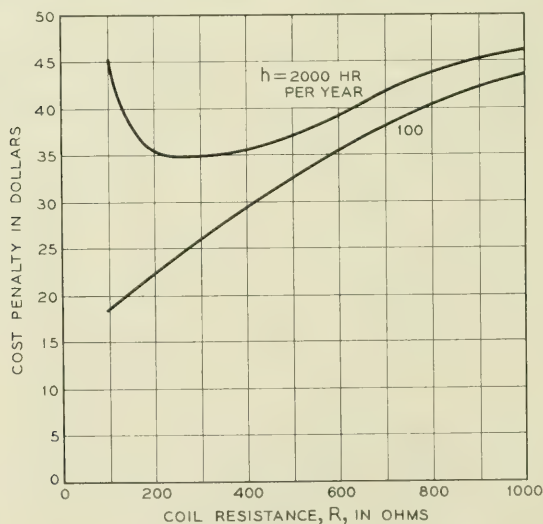


Fig. 16 — Best coil design for both speed and power economy.

most economical, over-all. The problem will first be discussed in a generalized form, and the means for getting the economy will then be reviewed.

6.1 GENERAL DISCUSSION OF CODING

If enough information were available to the circuit designer, he should be able to choose a relay for any particular application by considering

(a) The penalties in performance (expressed in terms of first cost) resulting from having only a certain limited number of available combinations of relays, as against complete flexibility, i.e., the penalties due to standardizing, and

(b) The penalties in first cost resulting from many variations of a basic type, as compared with one single standard type, i.e., the penalties due to *not* standardizing.

By weighing both the penalties and the advantages of standardization in each case, they may be maintained in approximate balance, and give the economical number of codes.

In the development of an entire new switching system using relays, there will be approximate ideas on the number of relays needed in the system, and on the number of kinds of relays required to fairly well satisfy the circuit functions. Based on past experience, the statistical distribution of kinds of relays for the various uses in relation to annual demand may also be approximated. The number of codes provided will determine the number of cells into which the statistical distribution is divided, and also the relative annual demand for each. In the extreme, only one code might be provided. It would have the advantage of very large lot-size, but also it would have to do the hardest as well as the easiest function. Such a requirement could lead to such absurdities as requiring that all relays have three windings, twelve transfers, best grade of contact metal, and the like; or to a greatly increased number of simpler relays. In addition, with but little flexibility in the design, there would be the added performance disadvantages due to imperfectly matching certain needs for power economy, speed, and the like. Hence, there is a large performance disadvantage, if but one kind of relay were provided.

At the other extreme would be cell divisions of kinds of relays to give the perfect match for every circuit use. Then, with no spare contacts, the best possible power economy, and every other condition at its ideal value, no performance penalty at all would be entailed. But every code increase would further subdivide the manufacture, till the lot-size penalties grew excessive.

These variations are shown by the curves in Fig. 1, and it is their sum,

as seen in the curve marked *c*, that should decide where best to work. In planning a whole new central office which may use from 40,000 to 80,000 relays, and a new design of relay whose manufacturing costs are not yet well established, the data to give numerical values for Fig. 1 can only be quite tentative, and the results can only serve as guides. Yet such work has been most useful in recent effort on a redesigned No. 5 crossbar system to use the wire spring relay. Calculations could be made with enough assurance to determine the general shape of the total penalty curve (Curve *c*), and showed an optimum value for kinds of relays of about 200. Even more important was the indication that the curve was extremely flat in its optimum region, giving but a small change in the total penalty if an error, even as great as 2 to 1, were made in the number of codes to be used.

6.2 THE DETAILED PROCEDURES OF CODE SELECTION

Background information as gained above helps in guiding the early steps in picking codes of relays for the job, but there is a better method of picking codes when the circuit designer is actually down to cases. By following the steps below, the optimum number of codes will automatically result.

The first step in the coding program is to pick a list of basic codes. This can be done quite arbitrarily, initially, and should be based on general knowledge of types of circuit functions, together with specific knowledge of certain uses with very large demands. For example, certain coils will be needed to cover the range from very fast operation with no concern for current drain, to slow release and emphasis on economy of power; there will be single and multiple windings as dictated by circuit functions. Also, certain arrangements of contacts may be assumed to span the range of needs from a single make up to the capacity of the design — twelve transfers or twenty-four makes in the case of the new wire spring relay. Further, certain combinations of springs and coils will be evident at once as ideally suited to particular large-demand uses, and these should be on the original list. Soon the likely minimum demand for each will be quite evident. Such a list of tentative codes and demands is the first step in the coding scheme.

Now, as each circuit application arises, this basic list may be consulted. For some cases, an available code will be ideal, but for many others a new arrangement may be desired. In each such case a simple calculation will show the economical choice between a new code or the old code with more features than are really needed. Steps in the calculations are given

below, where the extension of the old code is identified as Plan A, and the use of a new code as Plan B.

For Plan A, there will be a new demand, $(X + Y)$, which is greater than its previous demand by Y , the quantity of the new application. The demand cost penalty can then be read from charts such as Fig. 4. Performance penalties for pertinent features such as speed, power, extra contacts, etc. may be read from charts of the kind described in the earlier sections above. The sum of these values, multiplied by Y , is the cost penalty for the new application of an existing code.

For Plan B, there will be a lot-size cost penalty due to its demand. But there will also be a penalty imposed on the original code, (whose demand was X), because its demand was not enlarged to value $X + Y$. Thus the total lot-size cost penalty is made up of two parts:

(a) Penalty due to demand Y , multiplied by Y .

(b) [Penalty due to demand X - Penalty due to demand $(X + Y)$] multiplied by X .

The unit lot-size cost penalty for Plan B is the sum of these two factors, divided by the quantity Y . The design of relay for Plan B may be chosen by the methods previously outlined to be as nearly optimum as is feasible, and then performance penalties itemized as in Plan A above. The sum of these penalties, added to the lot-size penalties just mentioned, give a number for comparison with Plan A.

After the penalties due to either plan are compared, the least costly should be chosen. In cases where the costs are approximately equal, preference should probably be given to Plan A, as encouraging standardization.

In summary, the ideal number of codes in a newly designed system

TABLE IV — CHECK LIST FOR RELAY CODE SELECTION
AMOUNT OF EQUIVALENT COST PENALTY

Kind of Penalty	Plan A	Plan B
Lot-size (a)	Note (1)	Note (2)
(b)	$[(2) - (1)] \times \frac{X}{Y} =$	
Power		
Speed		
Extra parts		
Other		
Total		

Note (1): Fill in for demand $X + Y$.

Note (2): Fill in for demand X .

Note (3): Fill in for demand Y .

may be approximated by choosing the more economical alternative resulting from filling in the check list given in Table IV.

PART VII — SUMMARY

In the telephone central office some 60,000 relays are needed to carry out the automatic switching functions. Since the entire office is built around them, they determine the cost of the office, not only by their first cost, and maintenance cost, but at least equally importantly by their influence on the power plant, and the amount of common control equipment that is needed. There is opportunity for great economy, overall, when manufacturing cost is put in proper balance with the cost of power, cost of functioning time, and similar performance variables.

The preceding pages have shown how each factor in the total system cost may be stated in a common language — the equivalent first cost value. When each variable, expressed as an incremental figure denoting the cost compared to an ideal standard, is so stated, all relay factors in the switching system may be cross-compared. This permits a large number of optimizing procedures to be carried out.

Optimizing methods for several important cases have been illustrated, and relations or procedures of value to the relay designer are presented. Particularly outstanding cases are the means for realizing (a) power-plus-coil economy and (b) power-plus-speed economy.

On the side of manufacturing cost, relay designers need some guidance as to how their applications problems will be influenced by the lot-size, or annual demand. Optimizing procedures are also given for this problem, to yield approximate values for optimum lot-sizes under various manufacturing conditions, and the corresponding cost penalty as a function of annual demand.

The steps to be taken in order to strike the best balance between standardizing for maximum manufacturing economy and diversifying for maximum performance economy are also given.

The problem as a whole is a striking example of how design for service can be applied on a very large scale. The economies realized by the approach used here have avoided the expenditure of many additional thousands of dollars yearly to the telephone companies, with corresponding economy to the customer.

ACKNOWLEDGMENT

Appreciation is here expressed to R. L. Peek, Jr., and Mrs. K. R. Randall for helpful technical suggestions, and also to Mrs. Randall for help with the preparation of figures and examples.

Symbols

(Manuscript received October 20, 1953)

The following list gives the symbols used through the several articles appearing in this issue of the JOURNAL. The list does not include all the variant forms of the different symbols distinguished by subscripts, as these distinctions are indicated in the context where they are used.

MECHANICAL:

a	Cross sectional area
ℓ	Length
m	Effective armature mass
T	Kinetic energy
V	Work done to overcome static load
x	Armature displacement

ELECTRICAL:

c	Copper efficiency or fraction of the copper volume occupied by conductor
E	Applied battery voltage
G	Total equivalent single turn conductance
G_c	Equivalent single turn conductance of coil; N^2/R
G_E	Equivalent single turn eddy current conductance
G'_E	Effective single turn eddy current conductance; $G_E e^{-G_E/(G_c+G_s)}$
G_s	Equivalent single turn conductance of sleeve
i	Instantaneous current
I	Steady state current
m	Mean length of turn in winding
N	Number of turns in winding
L_1	Single turn inductance; $\frac{4\pi}{9\ell'(x)}$
NI	Ampere turns
NI_0	Just operate or just release ampere turn value
q	Ratio of just operate to final ampere turns; $\frac{NI_0}{NI}$

ρ	Resistivity of material
R	Resistance
S	Coil volume
v	Ratio of flux attained at time t to steady state flux
W	Power; I^2R

MAGNETIC:

A	Equivalent pole face area
A_2	Effective pole face area (design value)
B	Induction (flux density)
B'	Value of B for maximum permeability
B''	Saturation density
B_M	Density at $\mu = 1000$
B_R	Remanence (density at $H = 0$)
C_L	$\frac{\mathfrak{R}_0 + \mathfrak{R}_L}{\mathfrak{R}_0}$
F	Magnetic pull
$\mathfrak{F}$	Magnetomotive force; $4\pi NI$
$\mathfrak{F}_C$	Coercive magnetomotive force
H	Field intensity
H_C	Coercive force
L	Inductance
μ	Permeability
μ_A	Permeability of air
μ'	Maximum value of permeability
$\mathfrak{R}$	Reluctance
$\mathfrak{R}_C$	Reluctance of the core
$\mathfrak{R}'_C$	Minimum value of core reluctance
$\mathfrak{R}_1$	$\mathfrak{R}(x)$ for $x = x_1$
$\mathfrak{R}_i$	Initial incremental reluctance
$\mathfrak{R}_0$	Equivalent closed gap reluctance
$\mathfrak{R}_L$	Equivalent leakage reluctance
$\mathfrak{R}_{L2}$	Effective leakage reluctance (design value)
$\mathfrak{R}_{02}$	Effective closed gap reluctance (design value)
$\mathfrak{R}(x)$	Equivalent relay reluctance $\frac{\mathfrak{R}_L \left(\mathfrak{R}_0 + \frac{x}{A} \right)}{\mathfrak{R}_0 + \mathfrak{R}_L + \frac{x}{A}}$
φ	flux
Φ	Steady state flux

φ_1	Initial equilibrium flux
φ'	Flux for maximum permeability or minimum reluctance
φ''	Flux at saturation
φ_0	Residual flux
φ_G	Gap flux
U	Field energy
u	$x/A\mathcal{R}_0$ or x'/x_0
W	Mechanical work done by magnet
x_0	$A\mathcal{R}_0$

TIME:

t	Time
t_0	Operate or release time
t_1	Waiting time
t_2	Motion time
t_3	Stagger time
t_E	Eddy current time constant; $L_1 G'_E$
t_C	Winding time constant; $L_1 G_C$
t_S	Sleeve time constant; $L_1 G_S$

Abstracts of Bell System Technical Papers* Not Published in this Journal

BENNETT, W.¹

Telephone-System Applications of Recorded Machine Announcements, A.I.E.E., Trans., Commun. & Electronics **8**, pp. 478-483, Sept., 1953.

BIELING, D., see D. EDELSON.

BIGGS, B. S.¹ and W. L. HAWKINS¹

Oxidative Aging of Polyethylene, Modern Plastics, **31**, pp. 121-122, 124+, Sept., 1953 (Monograph 2155).

Thermal oxidation of polyethylene follows the pattern set by lower homologues such as paraffinic waxes and oils. It is an autocatalytic free radical chain reaction and is subject to inhibition by typical antioxidants. The rate of degradation in the dark at room temperatures is found to be extremely low. Photo-oxidation of polyethylene is rapid in contrast with that of saturated low molecular weight aliphatic hydrocarbons. Furthermore, antioxidants are of little benefit in protecting against exposure to light. Opaque pigments are of great value in reducing the effects of light, finely divided carbon black being particularly effective. By proper compounding polyethylene can be made to last many years outdoors.

BOZORTH, R. M., see H. J. WILLIAMS.

BROWN, W. L.¹

n-Type Surface Conductivity on p-Type Germanium, Phys. Rev., **91**, pp. 518-527, Aug. 1, 1953 (Monograph 2173).

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ Bell Telephone Laboratories.

A positive charge on the surface of a *p*-type germanium crystal induces a net negative space charge within the crystal adjacent to the surface. This space charge is composed of ionized acceptor atoms and also of electrons under certain conditions. When electrons occur they provide a layer of *n*-type conductivity immediately under the *p*-type germanium surface. Such a layer has been found on the *p*-type region of some *n-p-n* transistors. In the *n-p-n* structure the layer of electrons appears as an extra conducting path—"a channel"—across the *p*-type material between the two *n*-type ends. The conductance of a channel and the capacity between the channel and the *p*-type material have been measured and compared with the theoretical predictions based on a simple model.

CORNELL, L. P., see E. P. SMITH.

CRANE, G. R.⁴, F. HAUSER⁴ AND H. A. MANLEY⁴

Westrex Film Editor, J.S.M.P.T.E., **61**, Part 1, pp. 316-323, Sept., 1953.

This paper describes a film-editing machine which employs continuous projection resulting in quiet operation. It accommodates standard-picture and photographic or magnetic sound film as well as composite sound-picture film. Differential synchronizing of sound and picture while running, automatic fast stop and simplified threading features in the film gates with finger-tip release materially increase operating efficiency.

CRISSMAN, L.³

See Both Sides of the Game, Tele-Tech., **12**, p. 84, Sept., 1953.

DACEY, G. C.¹ AND I. M. ROSS¹

Unipolar "Field-Effect" Transistor, I.R.E., Proc., **41**, pp. 970-979, Aug., 1953.

Unipolar "field-effect" transistors of a type suggested by W. Shockley have been constructed and tested. The idealized theory of Shockley has been extended to cover the actual geometries involved, and design nomographs are presented. It is found that these structures can be designed in such a way as to yield a negative resistance at the input terminals. The characteristics of several units are presented and analyzed. It is shown that these characteristics are in substantial agreement with the extended theory. Finally a speculative evaluation of the possible future applications of field effect transistors is made.

¹ Bell Telephone Laboratories.

³ Western Electric Company.

⁴ Westrex Corporation.

EDELSON, D.¹, C. A. BIELING¹ AND G. T. KOHMAN¹

Electrical Decomposition of Sulfur Hexafluoride, Ind. and Eng. Chem., **45**, pp. 2094-2096, Sept., 1953.

EHRBAR, R. D., see C. H. ELMENDORF.

ELMENDORF, C. H.¹, R. D. EHRBAR¹, R. H. KLIE⁷, AND A. J. GROSSMAN¹

The L3 Coaxial System, A.I.E.E., Trans. Commun. & Electronics, **8**, pp. 395-413, Sept., 1953.

The L3 coaxial system is a new broad-band facility for use with existing and new coaxial cables. It makes possible the transmission of 1,860 telephone channels or 600 telephone channels and a television channel in each direction on a pair of coaxial tubes. The principal system design problems and the methods used in their solution are described in terms of its components and their location in the system.

FINE, M. E.¹

C_p-C_v in Silicon and Germanium, Letter to the Editor, J. Chem. Phys., **21**, P. 1427, Aug., 1953.

FRAYNE, J. G.⁴ AND E. W. TEMPLIN⁴

Stereophonic Recording and Reproducing Equipment, J.S.M.P.T.E., **61**, Part 2, pp. 395-407, Sept., 1953.

This paper describes new stereophonic recording channel equipment including a six-position mixer and portable three-channel recorder. For re-recording, the previously described triple-track recorder-reproducer is available. For review-room and theater reproduction, a theater-type dummy equipped for three-channel stereophonic reproduction is described.

GOERTZ, M., see H. J. WILLIAMS.

GRAMELS, J.¹

Problems to Consider in Applying Selenium Rectifiers, A.I.E.E., Trans., Commun. & Electronics, **8**, pp. 488-492, Sept., 1953.

GRAY, A. N.³

Room-Temperature Compound Process, Mech. Eng., **75**, pp. 625-628, Aug., 1953.

¹ Bell Telephone Laboratories.

³ Western Electric Company.

⁴ Westrex Corporation.

⁷ Sandia Corporation.

GROSSMAN, A. J., see C. H. ELMENDORF.

HAGSTRUM, H. D.¹

Electron Ejection From Ta by He^+ , He^{++} , and He_2^+ , Phys. Rev., **91**, pp. 543-551, Aug. 1, 1953 (Monograph 2148).

Measurements of total yield (γ_i) and kinetic energy distribution are reported for electrons ejected from tantalum by the ions, He^+ , He^{++} and He_2^+ in the kinetic energy range 10 to 1000 ev. The evidence presented indicates that the electrons are released by a collision of the second kind of the ion with the metal surface (potential ejection). One internal secondary electron is produced per incident ion. The probability of this electron escaping is reduced by the possibility of internal reflection at the image barrier at the metal surface. γ_i for the slowest ions is observed to be 0.14, 0.52, and 0.10 for He^+ , He^{++} , and He_2^+ , respectively. The data presented must be considered representative of gas-covered tantalum, since no gas was observed to desorb from the target on heating to 175°K. γ_i was found not to vary with time after cooling the target indicating rapid re-establishment of the equilibrium gas layer on the surface from within the metal. The work function of the covered Ta surface is found to be ca 4.9 ev, some 0.8 ev higher than that of atomically clean Ta. Considerations based on a theory which includes variation of energy levels near the metal surface show resonance neutralization of He^+ at the covered Ta surface not to be possible. Thus only the so-called direct process of potential ejection occurs, with which conclusion the measured energy limits of ejected electrons are in agreement.

HAUSER, F., see G. R. CRANE

HAWKINS, W. L., see B. S. BIGGS.

HEFFNER, H.¹

Backward-Wave Tube, Electronics, **26**, pp. 135-137, Oct., 1953.

Electron-stream amplifier utilizing backward-wave mode forms microwave oscillator continuously tunable over a three-to-one bandwidth by a single voltage control. Tubes have been built for frequency centering about 6,000, 10,000 and 50,000 mc.

HENDERSON, O.⁵

Magnetic Amplifier Controls for Rectifier Protecting Underground Metallic Structures Cathodically, Corrosion, **9**, pp. 216-220, July, 1953.

This paper covers the initial attempt on the part of the Ohio Bell Telephone

¹ Bell Telephone Laboratories.

⁵ Ohio Bell Telephone.

Company to use a magnetic amplifier to give continuous control of the output of a copper-oxide rectifier used for cathodic protection of underground lead covered cables. The controlled rectifier was designed for use at a location where stray current "end effects" were damaging underground telephone cables and municipal light cables and were threatening high-pressure water mains. When properly adjusted, the output of the rectifier will increase or decrease automatically so that at each instant the amount of forced drainage will be adequate to protect the underground telephone cable sheath but will not be in excess of the value at which neighboring underground metallic structures would become anodic and would thus become subject to corrosion. For satisfactory operation the amplifier had to be designed to give a large gain so that a change in the control voltage of only 0.2 volt (+0.1 to -0.1 volt) would be sufficient to vary the output of the rectifier from practically zero current to full rating. In the installation described in this paper the gain of the magnetic amplifier is in the order of 12,000. The magnetic amplifier type of control is well suited to outdoor installations subject to wide changes in temperature and at remote locations where frequent maintenance inspections are not feasible. The magnetic amplifier has advantages over other available control devices which accomplish this same purpose in that it has no moving parts, no vacuum tubes, batteries, motors, relays or contactors. The magnetic amplifier gives a continuous output control which is superior to the stepped increment changes that result from relay or contactor operation.

JERONE, M. G., see E. P. SMITH.

JOHNSON, J. B.¹ AND K. G. MCKAY¹

Secondary Electron Emission of Crystalline MgO, Phys. Rev., **91**, pp. 582-587, Aug. 1, 1953 (Monograph 2163).

Secondary emission is measured from single crystals of MgO cleaved along the (100) plane. The maximum ratio of secondary to primary current, $\delta_{\max}$, is about 7 at about 1,000 volts and room temperatures. The cross-overs are at 33 volts and far above 5,000 volts. Most probable energy of emission is 1 ev or less. A definite effect of temperature is established, decreasing with increasing temperature, in accord with expectations for an insulator.

JONES, T. A.¹ AND W. A. PHELPS¹

Level Compensator for Telephotograph Systems, Elec. Eng., **72**, pp. 787-791, Sept., 1953.

To eliminate interference in telephotograph transmission through broadband carrier equipment, it was decided to cancel it from the signal delivered by the carrier facility instead of modifying the carrier equipment. Conse-

¹ Bell Telephone Laboratories.

quently, a recently developed telephotograph level compensator, consisting of a pilot channel arrangement designed for insertion in the telephotograph connecting circuits, is utilized.

KISTLER, R. E.⁶

Radio Links U.S. and Canada, *Telephony*, **145**, pp. 16-17, 33, Sept. 5, 1953.

KLIE, R. H., see C. H. ELMENDORF.

KOHMAN, G. T., see D. EDELSON.

KRUSE, P. F., Jr.⁷ and W. B. WALLACE⁷

Identification of Polymeric Materials, *Anal. Chem.*, **25**, p. 1156, Aug., 1953.

KUH, E. S.¹

Potential Analog-Network Synthesis for Arbitrary Loss Functions, *J. App. Phys.*, **24**, pp. 897-902, July 1953.

A general method is developed for designing networks with arbitrary loss functions based on the potential analogy. An appropriate potential problem is formed on the basis of the given loss function by introducing continuous charge distribution on the complex frequency plane. After the potential problem is solved, the technique of quantization of charge is used to find the natural modes of the network function.

LANGE, R. W.¹

40- to 4,000-Microwatt Power Meter, *A.I.E.E. Trans., Commun. & Electronics*, **8**, pp. 492-494, Sept., 1953.

LEWIS W. D.¹

Electronic Computers and Telephone Switching, *I.R.E., Proc.*, **41**, pp. 1242-1244, Oct., 1953.

Automatic telephone switching and digital computation have much in common. Both rely upon discrete rather than continuous devices. Development of recent switching systems with a close functional resemblance to large digital computers has increased this overlap. The next big step in

¹ Bell Telephone Laboratories.

⁶ Pacific Telephone and Telegraph Company.

⁷ Sandia Corporation.

telephone switching should be towards electronics. In making this step, switching scientists and engineers can be much helped by modern electronic computer technology. To be successful, they must also contribute to this technology.

LINDHOM, P.³

"Feedback" — Heart of Automatic Process Control, *Factory*, 111, pp. 106–109, Oct., 1953.

LOGAN, R. A.¹

Thermally Induced Acceptors in Single Crystal Germanium, Letter to the Editor, *Phys. Rev.*, 91, pp. 757–758, Aug. 1, 1953.

MAITA, J. P., see M. TANENBAUM.

MALLERY, P.¹

Transistors and Their Circuits in the 4A Toll Crossbar Switching System, *A.I.E.E. Trans., Commun. & Electronics*, 8, pp. 388–392, Sept., 1953.

MALTHANER, W. A.¹ AND H. E. VAUGHAN¹

Automatic Telephone System Employing Magnetic Drum Memory, *I.R.E., Proc.*, 41, pp. 1341–1347, Oct., 1953 (Monograph 2151).

The use of magnetic drum memory in an automatic telephone switching office is described. A capacitive scanner acts as a time-division connector through which information generated by subscribers' telephone sets is conveyed to storage on magnetic drums. Information thus accumulated is combined with "permanent" information on the magnetic drums, processed in accordance with built-in programs and dispatched to control call switching circuits. Technical feasibility of this system has been demonstrated by the construction and successful operation of a large-scale laboratory model.

MANLEY, H. A., see G. R. CRANE.

MATTHIAS, B. T.¹

Superconducting Compounds, Letter to the Editor, *Phys. Rev.*, 91, p. 413, July 15, 1953.

McAFEE, K. B., see K. G. McKAY.

¹ Bell Telephone Laboratories.

³ Western Electric Company.

McCARTHY, R. H.³

Organization for Production Engineering, Mech. Eng., **75**, pp. 785–788, 793, Oct., 1953.

McGUIGAN, J. H.¹

Combined Reading and Writing on a Magnetic Drum, I.R.E., Proc., **41**, pp. 1438–1444, Oct., 1953 (Monograph 2152).

This paper points out that the characteristics of magnetic recording make it possible to combine reading and writing in the same cell as it passes just once under the head. Amplifier requirements for this method of operation are discussed and a suitable design presented. A single head is used for both reading and writing. The process can be repeated in every successive cell at a cell rate of 60 kc. The techniques described, which are applicable to either parallel or serial systems, extend the utility of magnetic drums by allowing data processing as well as data storage.

McKAY, K. G., see J. B. JOHNSON.

McKAY, K. G.¹ AND K. B. McAFEE¹

Electron Multiplication in Silicon and Germanium, Phys. Rev., **91**, pp. 1079–1084, Sept. 1, 1953 (Monograph 2162).

Electron multiplication in silicon and germanium has been studied in the high fields of wide *p-n* junctions for voltages in the pre-breakdown region. Multiplication factors as high as eighteen have been observed at room temperature. Carriers injected by light, alpha particles, or thermal-generation are multiplied in the same manner. The time required for the multiplication process is less than 2×10^{-8} second. Approximately equal multiplication factors are obtained for injected electrons and injected holes. The multiplication increases rapidly as "breakdown voltage" is approached. The data are well represented by ionization rates computed by conventional avalanche theory. In very narrow junctions, no observable multiplication occurs before Zener emission sets in, as previously reported. It is incidentally determined that the efficiency of ionization by alpha particles bombarding silicon is 3.6 ± 0.3 electron volts per electron-hole pair produced.

McSKIMIN, H. J.¹

Measurement of Elastic Constants at Low Temperatures by Means of Ultrasonic Waves — Data for Silicon and Germanium Single Crystals and for Fused Silica, J. App. Phys., **24**, pp. 988–997, Aug., 1953 (Monograph 2171).

¹ Bell Telephone Laboratories.

³ Western Electric Company.

Ultrasonic waves (shear on longitudinal) in the 10-30 mc range are transmitted down a fused silica rod, through a polystyrene or silicone one-quarter wavelength seal, and into the solid specimen. Measurement of reflections within the specimen yields values for velocities of propagation and elastic constants. Data obtained over a temperature range of 78° to 300°K for silicon and germanium single crystals, and 1.6° to 300°K for fused silica are listed. For the latter, a high loss is noted, with an indicated maximum near 30°K.

MERZ, W. J.¹

Double Hysteresis Loop of BaTiO₃ at the Curie Point, Phys. Rev., **91**, pp. 513-517, Aug. 1, 1953 (Monograph 2166).

It is known that the Curie point of the ferroelectric BaTiO₃ shifts to higher temperatures when a dc bias field is applied. If the crystal shows a sharp transition, we expect by applying an ac field at the Curie temperature that the crystal would become alternately ferroelectric and nonferroelectric in the cycle of the ac field. This can be seen in the shape of the hysteresis loop at temperatures slightly above θ . In the center of the polarization P versus field E plot, we observe a linear behavior corresponding to the paraelectric state of BaTiO₃ above θ . At both high voltage ends, however, we observe a hysteresis loop corresponding to the ferroelectric state. A change in temperature causes a change in size and shape of the double hysteresis loops, ranging from a line with curves at the ends (higher temperature) to two overlapping loops (lower temperature). The results obtained allow us to calculate the different constants in the free-energy expression of Devonshire and Slater. One of the results shows that the transition is of the first order since the P^4 term turns out to be negative. The properties of the hysteresis loops are discussed, especially the large spontaneous electrical polarization and the low coercive field strength.

MOORE, E. F., see C. E. SHANNON.

MORRISON, J.¹

Leak Control Tube., Rev. Sci. Instr., **24**, pp. 564-547, July, 1953.

PHELPS, W. A., see T. A. JONES.

PRINCE, M. B.¹

Experimental Confirmation of Relation Between Pulse Drift Mobility and Charge Carrier Drift Mobility in Germanium, Phys. Rev., **91**, pp. 271-272, July 15, 1953 (Monograph 2168).

Experimental data of drift mobilities of minority carriers in germanium are

¹ Bell Telephone Laboratories.

brought into agreement with theoretical predictions by distinguishing between group velocity and particle velocity of a pulse of minority carriers. Corrected high temperature measurements of electron drift mobility are consistent with the theoretical prediction $\mu = AT^{3/2}$. The experimentally determined value of A is $2.0 \times 10^7 \text{ cm}^2 \text{ deg}^{3/2}/\text{volt-sec}$.

REISS, H.¹

Chemical Effects Due to the Ionization of Impurities in Semiconductors, J. Chem. Phys., **21**, p. 1209, July, 1953 (Monograph 2172).

This paper contains a theoretical account of the possible effects which the ionization of donor and acceptor impurities can induce in the thermodynamic phase relations involving their solutions with semiconductors. These effects reflect the energy band structure of the semiconductor. Furthermore, the phenomenon has an interest of its own, for within a certain range of experimental conditions the effects can be attributed to a chemical-like, mass action behavior of the electrons which play the roles of negative ions. Section V is a brief discussion of a fine point concerning the Fermi level. It is shown that although the Fermi level is certainly the electronic electrochemical potential, it is not the Gibbs free energy per electron unless the density of electron energy levels is linear in the volume of the system.

ROBBINS, R. L.¹

Measurement of Path Loss Between Miami and Key West at 3675 mc., I.R.E., Trans., P.G.A.P. **1**, pp. 5-8, July, 1953.

Radio transmission measurements have been made at 3.675 megacycles on the 130-mile path between Miami and Key West, Florida, which is largely over watery wastes of the Everglades and shallow sea waters of the Florida Keys. Path loss and fading characteristics for this terrain were not found to differ materially from the characteristics of hilly or mountainous paths in the northeastern section of the country.

ROSS, I. M., see G. C. DACEY.

SHANNON, C. E.

Turing's Formulation of Computing Machines and Von Neumann's Models of Self-reproducing Machines., I.R.E., Proc., **41**, pp. 1235-1241, Oct., 1953 (Monograph 2150).

This paper reviews briefly some of the recent developments in the field of automata and non-numerical computation. A number of typical machines are described, including logic machines, game-playing machines and learning machines. Some theoretical questions and developments are discussed, such as a comparison of computers and the brain.

¹ Bell Telephone Laboratories.

SHANNON, C. E.¹ AND E. F. MOORE¹**Machine Aid for Switching Circuit Design**, I.R.E., Proc., **41**, pp. 1348-1351, Oct., 1953 (Monograph 2153).

Design of circuits composed of logical elements may be facilitated by auxiliary machines. This paper describes one such machine, made of relays, selector switches, gas diodes, and germanium diodes. This machine (called the relay circuit analyzer) has as inputs both a relay contact circuit and the specifications the circuit is expected to satisfy. The analyzer (1) verifies whether the circuit satisfies the specifications, (2) makes systematic attempts to simplify the circuit by removing redundant contacts, and also (3) obtains mathematically rigorous lower bounds for the numbers and types of contacts needed to satisfy the specifications. A special feature of the analyzer is its ability to take advantage of circuit specifications which are incompletely stated. The auxiliary machine method of doing these and similar operations is compared with the method of coding them on a general-purpose digital computer.

SMITH, E. P.⁶, L. P. CORNELL⁶ AND M. G. JEROME⁶**Co-ordinating M1 and N1 Telephone Carrier Systems**, Elec. Eng., **72**, p. 780, Sept., 1953.TANENBAUM, M.¹ AND J. P. MAITA¹**Hall Effect and Conductivity of InSb Single Crystals**, Letter to the Editor, Phys. Rev., **91**, pp. 1009-1010, Aug. 15, 1953.

TEMPLIN, E. W., see J. G. FRAYNE.

TOWSLEY, L. M., see E. A. WOOD.

TUCKER, C. J., JR.⁸**Emergency Reporting System Installed by Southern Bell**, Telephony, **145**, pp. 26, 38, Aug. 29, 1953.

New fire and emergency reporting system utilizing telephone facilities for the general public to make verbal reports of fires and other emergencies — the first of its kind in the nation — was put into service in Miami, Fla., on Aug. 1, by Southern Bell Telephone & Telegraph Co.

VAN ROOSBROECK, W.¹¹ Bell Telephone Laboratories.⁶ Pacific Telephone and Telegraph Company.⁸ Southern Bell Telephone Company.

Transport of Added Current Carriers in a Homogeneous Semiconductor, Phys. Rev., **91**, pp. 282-289, July 15, 1953 (Monograph 2165).

Taking into account the thermal equilibrium minority carrier concentration and employing the formulation which includes, as one of two fundamental equations, the continuity equation for added carrier concentration Δp , this equation is derived in a form which exhibits the ambipolar nature of the diffusion, drift, and recombination mechanisms under electrical neutrality. The general concentration-dependent diffusivity is given. The local drift velocity of Δp has the direction of total current density in an n -type semiconductor and the reverse in a p -type semiconductor, differing in general in both magnitude and direction from the minority-carrier drift velocity. Specifying a model for recombination fixes the dependence of a lifetime function for Δp on Δp and the electron and hole mean lifetimes. Negative Δp , or carrier depletion with electrical neutrality, may occur. For known total current density, the continuity equation alone suffices, as for the case of $|\Delta p|$ small, for which the equation is linear. A condition for this comparatively important case is derived, and theoretical relationships are given with the aid of a parameter specifying the Fermi level, which determine for germanium the minority carrier- Δp drift velocity ratio as well as the ambipolar diffusivity and group mobility in terms of resistivity and temperature.

VAUGHAN, H. E., see W. A. MALTHANER.

VOGELSONG, J. H.¹

Transistor Pulse Amplifier Using External Regeneration, I.R.E., Proc., **41**, pp. 1444-1450, Oct., 1953.

Pulse-regenerative amplifier using a point-contact transistor has been operated at a basic frequency of 3 megacycles. To produce regenerated pulses with waveshapes which are practically independent of the waveshapes of the input pulses, a germanium diode circuit has been used in conjunction with an external feed-back path. This arrangement also provides for the synchronization of the output pulses with a master clock. Transformer coupling has been incorporated into the circuit to provide dc restoration.

WALKER, A. C.¹

Hydrothermal Synthesis of Quartz Crystals, Am. Ceramic Soc., J., **36**, pp. 250-256, Aug., 1953 (Monograph 2146).

Research at the Bell Telephone Laboratories on the problem of growing large single crystals of quartz has now progressed to a point where it is possible to grow crystals weighing more than 1 lb. each in a period of 60 days or less. Equipment now in use includes autoclaves 4 inches in inside diameter and 4 ft. long, weighing about 1,150 lb. each. In developing the

¹ Bell Telephone Laboratories.

hydrothermal process used to grow these quartz crystals it has been necessary to solve many problems in the little-known field of high pressure. The results point to the possibility of growing other types of crystals, and the field of usefulness of this process now appears to be much more extensive than was the case at the beginning of the investigation in 1946. Of prime importance is the fact that crystals grown from solution are likely to be better formed and of more perfect quality than those grown from the melt or by other methods. Many of the difficulties inherent in this work have been due to corrosion of steel in alkaline solution. This was a rather unexpected problem, since in high pressure steam boilers alkali is added in small amounts to prevent corrosion of the boiler tubes. Such corrosion has been shown to be responsible for the appearance of electrical twinning on the growing faces of the quartz crystals. Other causes of such twinning have also been found in the course of this work.

WALKER, L. R.¹

Starting Currents in the Backward-Wave Oscillator, *J. App. Phys.*, **24**, pp. 854-859, July, 1953.

The starting current of a simple model of the backward-wave oscillator described by Kompfner and Williams has been calculated. The effect of space charge is included. The starting current I_0 may be written in the form

$$\frac{4V_0}{Z_0} \left[\frac{a_0(4QC)}{2N} \right]^3,$$

where V_0 is the beam voltage, Z_0 is the impedance of the circuit, N is the length of the oscillator in wavelengths measured on the circuit and $a_0(4QC)$ is a dimensionless quantity which has been evaluated as a function of the space-charge parameter $4QC$.

WALLACE, W. B., see P. F. KRUSE, JR.

WASHBURN, S. H.¹

Application of Boolean Algebra to the Design of Electronic Switching Circuits, *A.I.E.E. Trans., Commun. & Electronics*, **8**, pp. 380-388, Sept., 1953.

WERNER, D. R.²

Effects of Polarization on Telephone Cable Buried Through a Salt Bed, *Corrosion*, **9**, pp. 232-236, July, 1953.

Cathodic protection has been applied to a copper jacketed cable in a salt lake bed about one mile wide in which the earth resistivity was apparently uni-

¹ Bell Telephone Laboratories.

² American Telephone and Telegraph Company.

form at about 20 ohm-centimeters. Current of about one ampere flowed on the copper jacket into the low resistivity areas on either side and was found to be sustained by polarization effects when the cathodic protection current was removed. The current loss on the copper jacket was found to be concentrated in an area about 720 feet wide, 360 feet each side of the point where the cathodic protection current had been drained from the copper jacket. The copper jacket to soil potential tested most negative to a copper sulfate half cell in the 720-foot area where the current loss was concentrated and was of the order of -1.0 to -1.1 volts. Permanent remedial measures will consist of installations of magnesium anodes distributed throughout the low earth resistivity area and insulating joints in the copper jacket at locations where large changes in the earth resistivity occur.

WILLIAMS, H. J.¹, R. M. BOZORTH¹ AND M. GOERTZ¹

Mechanism of Transition in Magnetite at Low Temperatures, Phys. Rev., **91**, pp. 1107–1115, Sept. 1, 1953 (Monograph 2149).

When magnetite is cooled through -160°C it is known to undergo a transition (cubic to orthorhombic) that is influenced by the presence of a magnetic field. Our experiments are in agreement with the following mechanism of the transition: The orthorhombic c axis is parallel to one of the original cubic axes and is the axis of easiest magnetization. Generally, different regions of the original crystal will transform with their c axes lying along different cubic axes, and when no field is applied there are 6 different orientations which different regions assume. When a field is applied during a cooling a c axis tends to lie along the original cubic axis that is nearest to the applied field, the a and b axis having less but different tendencies to lie parallel to the field. Six magnetic crystal anisotropy constants are derived from torque curves measured in the (100) and (110) planes. From them magnetization curves are calculated for the 100 and 110 directions, and these are in agreement with experiment.

WOOD, E. A.¹ AND L. M. TOWSLEY¹

Manganese Film-Shield for FeK X-Rays, Rev. Sci. Instr., **24**, p. 547, July, 1953.

¹ Bell Telephone Laboratories.

Contributors to this Issue

ARTHUR C. KELLER. B.S., Cooper Union, 1923; M.S., Yale University, 1925; E.E., Cooper Union, 1926; Columbia University, 1926-30; Western Electric Company, 1917-25; Bell Telephone Laboratories, 1925-. Special Apparatus Development Engineer, 1943; Switching Apparatus Development Engineer, 1946; Assistant Director of Switching Apparatus Development, 1949; Director of Switching Apparatus Development, 1949-. Mr. Keller's experience in the Bell System includes development and design of telephone instruments; development of systems and apparatus for recording and reproducing sound; and, during World War II, the development, design, and preparation for manufacture of sonar systems and apparatus. His department, in addition to being responsible for a number of military projects, is responsible for the fundamental studies of switching apparatus and the development, design, and preparation for manufacture of electromagnetic and electromechanical switching apparatus for telephone systems. Member of the American Physical Society, A.I.E.E., Acoustical Society of America, I.R.E., S.M.P.T.E., and the Yale Engineering Association. Representative for Bell Telephone Laboratories in the Society for Experimental Stress Analysis. For his contributions to the Navy during World War II, he received awards from the Bureau of Ships and the Bureau of Ordnance.

MASON A. LOGAN, B.S. in Physics and Engineering, California Institute of Technology, 1927; M.A. in Physics, Columbia University, 1933; Bell Telephone Laboratories, 1927-. His early Laboratories' projects were concerned with wire transmission problems particularly those of circuits, noise and cross induction in local, manual and dial telephone circuits. This was followed by circuit research on alternating current methods of signaling including the use of non-linear elements and electronic terminal equipment. From 1941 to 1948 he worked on military projects, including a mine fire control system, anti-aircraft gun director, magnetic proximity fuses, and guided missiles. For the past five years he has been a member of the Switching Apparatus Development Department in which he is supervising a group concerned with static and dynamic behavior of

new electromagnets and relays. He is also engaged in investigations of the performance of electrical contacts on telephone relays.

ROBERT L. PEEK, JR., A.B. and Met.E., Columbia University, 1921 and 1923; Bell Telephone Laboratories, 1924-. In the Chemical Research Department and later the Apparatus Development Department, Mr. Peek's work related to the analytical and testing aspects of materials development. Since 1936 he has been engaged in apparatus development projects, including the wire spring relay and, during World War II, underwater ordnance and magnetostriction sonar.

H. N. WAGAR, B.S. in Physics, Harvard University, 1926; M.A. in Physics, Columbia University, 1931; Bell Telephone Laboratories, 1926-. Mr. Wagar has worked on the design of nearly all types of telephone relays as well as magnets, insulating methods and allied apparatus and practices. He was also associated with the preparation of text and presentation of training courses in this field. During World War II his projects for the military included work on an antiaircraft director and a proximity fuse for magnetic mines. He is currently Switching Apparatus Engineer, in charge of fundamental studies on electromagnetic switching apparatus, including contacts and magnetics.

THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII

MARCH 1954

NUMBER 2

Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks C. J. TRUITT 277

Intertoll Trunk Concentrating Equipment D. F. JOHNSTON 303

The Transistor as a Network Element J. T. BANGERT 329

Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath B. M. WOJCIECHOWSKI 353

The Application of Designed Experiments to the Card Translator C. B. BROWN and M. E. TERRY 369

Wave Propagation Along a Magnetically-Focused Cylindrical Electron Beam W. W. RIGROD and J. A. LEWIS 399

Diffraction of Plane Radio Waves by a Parabolic Cylinder S. O. RICE 417

Abstracts of Bell System Technical Papers Not Published in this Journal 505

Contributors to this Issue 514

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

- S. BRACKEN, *Chairman of the Board,*
Western Electric Company
- F. R. KAPPEL, *President, Western Electric Company*
- M. J. KELLY, *President, Bell Telephone Laboratories*
- E. J. McNEELY, *Vice President, American Telephone*
and Telegraph Company

EDITORIAL COMMITTEE

- E. I. GREEN, *Chairman*
- | | |
|---------------|---------------|
| A. J. BUSCH | F. R. LACK |
| W. H. DOHERTY | W. H. NUNN |
| G. D. EDWARDS | H. I. ROMNES |
| J. B. FISK | H. V. SCHMIDT |
| R. K. HONAMAN | G. N. THAYER |

EDITORIAL STAFF

- J. D. TEBO, *Editor*
- M. E. STRIEBY, *Managing Editor*
- R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXIII

MARCH 1954

NUMBER 2

Copyright, 1954, American Telephone and Telegraph Company

Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks

By C. J. TRUITT

(Manuscript received November 23, 1953)

In 1945 the Bell System embarked on an extensive study with the purpose of developing a program for operator toll dialing on a nationwide basis. Operator toll dialing had been done, of course, on a limited scale in various parts of the country for many years, but the concept of this program was one of nationwide proportions carried on with a uniform numbering plan arrangement and a completely integrated trunking system which would handle traffic at a high speed between any two points in the United States and Canada, even in the busier hours of the day.*

Implementation of this program required the development of new switching mechanisms and the exploitation of carrier transmission potentialities to a degree never before achieved. Great strides had already been made in these fields, resulting in the practical development of the coaxial cable system and the first toll crossbar switching office installed at Philadelphia in 1943. But the very core of the nationwide dialing plan was the proposal to revolutionize the method of traffic distribution so as to combine high speed handling over the intertoll trunk network with a highly efficient use of facilities. The method of accomplishing is called "engineered alternate routing"

* W. H. Nunn, Nationwide Numbering Plan, *Communication and Electronics*, **2**, Sept., 1952 and B. S. T. J., **31**, Sept., 1952.

in which a first choice direct route for traffic would be provided with trunks in such numbers as to force a predetermined portion of the load to seek another route where such residue or overflow would be carried at less cost per unit and with little or no delay. The principles of alternate routing and their application to interoffice trunking systems are the subject of this paper.

PRINCIPLES OF ALTERNATE ROUTING

Broadly speaking, alternate routing of telephone traffic is a procedure whereby a parcel of traffic is provided a second choice path to its destination which is also the second choice path for one or more other parcels of traffic. The use in common of this second choice path by the several parcels results in an overall requirement of trunk paths to the particular destination which is less than would be the case if each parcel had sole access to its own group of paths — and this is accomplished without impairment of the average speed with which all of the traffic is handled.

The familiar graded arrangements of trunk terminals in dial central office switching systems are examples of the trunking efficiency gained from alternate routing. It should be noted at once that the most efficient way to handle traffic from one central office to another is by means of a single group of trunks to which all traffic so destined has access. However, many such loads are so large that the number of trunks required to handle them on a single group often exceeds the practical terminal capacity of ordinary switching systems. Therefore, it becomes necessary to present portions of such a load, each to an individual subgroup. For example, assume that a busy-hour load of 334 calls of 100 seconds duration each (334 CCS) is to be carried at a probability of delay of one per cent (Poisson) from office M to office N and that the equipment permits access to a maximum of only ten trunk terminals for any one trunk group. Since the number of trunks required for the load in question is 17.5 (18) in a single group the establishment of three individual subgroups is indicated. To maintain P.01 service with three individual subgroups requires 8.3 trunks in each, assuming equal division of the load. The efficiency of each subgroup is 13.4 CCS per trunk which is some 30 per cent less than that of the larger single group which is 19.1 CCS.

Thanks to graded multiple arrangements,* such a heavy loss in efficiency need not be accepted. Pursuing our example, consider the load in question to be presented to an appropriate grade. This is illustrated schematically in Fig. 1.

* R. I. Wilkinson, The Interconnection of Telephone Systems — Graded Multiples, B. S. T. J., 10, Oct., 1931.

The capacity of this particular grade is precisely 334 CCS and its efficiency is 16.7 CCS per trunk or only 12 per cent below the single trunk group efficiency. Here the common trunks serve as an alternate route for such portions of the loads a, b, and c, respectively, as can not be handled by the five trunks which are individual to each. Because it is not likely that a, b, and c will overflow equal amounts of traffic at the same time, the common trunks are kept busy by a more or less complementary pattern of greater and lesser overflows at any given moment, emerging from the three subgroups. It is this action of the overflows, amply substantiated by experience, which accounts for the efficiency of graded multiples. Looked at another way, it may be said that all trunks above five in each subgroup of the split-multiple case have been pooled for the common use of all subgroups and that in so doing, it is possible to reduce the number of pooled trunks from 9.9 to five without

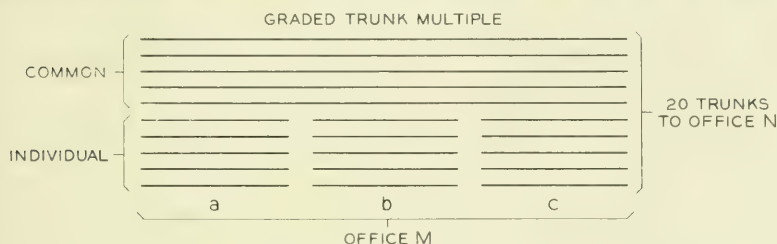


FIG. 1 — Typical graded arrangement of 20 trunks on 10 terminals.

impairing the speed of service. This very brief discussion of graded multiples serves merely to point out by familiar example, some of the potentialities of the alternate routing principle in the economical handling of telephone traffic.

ALTERNATE ROUTING IN LOCAL INTEROFFICE TRUNK NETWORKS

The effectiveness of alternate routing as illustrated by its action in graded multiples suggests the possibility of improving the efficiency of trunking between central offices by arranging the offices themselves in a sort of grade. Let us carry the analogy as far as practicable and assume that the loads a, b, and c in Fig. 1 are now emanating from central offices A, B, and C and still destined for office N. Let us assume that A, B, and C are typical offices in a multi-office city which has a tandem office, T, and further, that every office in the city has a group to and a group incoming from the tandem office. Fig. 2 illustrates these conditions with respect to offices A, B, C, and N and for simplicity indicates

groups carrying traffic in only one direction. The dashed lines represent the direct trunk groups between central offices and will be referred to as "high usage" groups. The solid lines represent trunk groups in the alternate route and will be referred to as "final" groups since they constitute the last choice path for traffic to N. It will be seen that the groups AN, BN, and CN correspond roughly to the groups of individual trunks serving a, b, and c in Fig. 1 and that group TN corresponds to the common trunks. But here the analogy breaks down, for the traffic from A, B, and C, respectively, to N will rarely be equal and there is no counterpart in Fig. 1 to the groups from A, B, and C to T. At this point it becomes clear that the techniques for determining the capacities of simple graded multiples are of no avail in determining the proper number of trunks in each group of such a system. The problem is further complicated by the fact that offices A, B, and C will offer to their respective tandem groups traffic for which such groups are the only route. Such traffic would consist of a number of relatively small items destined for other offices, not indicated in Fig. 2, and for which direct routing would be uneconomical.

To make such a system work there had to be available at the originating office a switching mechanism capable of testing two or more trunk groups in succession for a single call. The No. 1 local crossbar office developed in the 1930's provided such facilities and a trial of the alternate routing scheme for trunking between local offices was made at two New York City offices in 1941.

The basic aim was to so arrange the trunking layout at each office that the traffic to every other office in the city would be carried as economically as possible over some combination of available routes. This objective required determination of the cost of a direct route from the trial office to every other office and also the cost of any potential alternate route. The relationship between the costs of the several routes available to a given item, not the absolute values of the facilities was the important factor. It is not the intent here to discuss the derivation of such factors but rather to describe how they were used in determining the interoffice trunk layout.

The end result of the examination of line and switching costs for the several routes was a ratio of the cost of each potential alternate route to the cost of the direct route to each distant office. The smallest of such ratios would determine the alternate route to be used unless the ratio were unity or less, a rare circumstance which would indicate that no direct trunks should be provided. The next step in the problem was to provide trunks in such numbers and arrangements that the cost of

carrying the load from the originating office to each distant office would be at a minimum. This required some means of determining how much load would be carried by the direct trunks when offered a given load and consequently how much of that load would be overflowed to the alternate route. For this purpose, a formula* known as the Erlang B "lost-calls-cleared" assumption was used. This formula states for a given random offered load, the amount of load which will be carried by each of a number of trunks, n , tested in succession provided that the calls failing to be carried on the first attempt (the lost calls) *are not reoffered within the hour during which the first offering took place*. The condition italicized is extremely important to the problem since it requires that calls lost on the direct high usage group, i.e., the calls overflowed to the

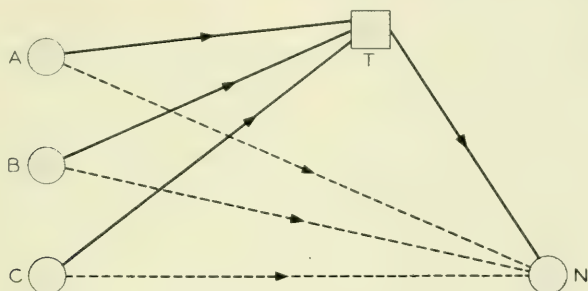


FIG. 2 — Illustration of simple interlocal trunk network arranged for alternate routing.

alternate route, must be disposed of without delay on the alternate route or routes. In the New York City trials it was assumed that the then current basis of provision of trunks in each leg of the alternate route (final groups AT and TN, Fig. 2) namely, with a probability of delay of one per cent. (P.01) would, as a practical matter, satisfy the condition that calls overflowing from AN should be cleared. It should be mentioned in passing that the results of the trials substantiated the reasonableness of this assumption.

A typical Erlang B distribution is shown in Fig. 3, Curve A wherein the load carried by each of $n=14$ trunks is shown for the condition of 240 offered CCS. Thus, assuming the load to be offered in succession to trunks 1, 2, 3, etc., in that order, it will be seen that the first trunk carries the most load, the second trunk somewhat less, the third still less until the fourteenth trunk carries about 0.5 per cent of the total. By

* A. K. Erlang, Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges, Post Office Electrical Engineers' Journal, 10, 1917.

taking the sum of the loads carried by a stated number of trunks and subtracting it from the total offered load, the amount of load "lost" or overflowed can be obtained. It should be noted here that the load carried by a given number of trunks under the pressure of a constant value of offered load, and provided lost calls are cleared, will always be the same whether or not the load is presented to said number of trunks in consecutive order. Thus, seven trunks to which 240 CCS are presented will carry 185 CCS (Fig. 3, Curve B) whereas six trunks will carry only 165 CCS, all values to the nearest whole CCS. Therefore, it may be said that the "last" trunk in a group of seven carries 20 CCS under the Erlang B concept when the offered load is 240 CCS, since by adding one trunk to a group of six, the capacity of the group to carry load is increased by 20 CCS. The significance of this notation will appear presently.

Let us leave the subject of direct high usage trunk group loading for a moment and consider the loading of final trunk groups such as AT and TN, in Fig. 2. We said earlier that an office such as A originates traffic of such small volume to each of many other offices in the city that direct routing would be economically unsound. Such items in the aggregate constitute a load of considerable size and all are routed via a tandem office for distribution to their respective destinations together with similar small parcels originating at B, C, etc. As stated earlier trunks to and from tandem were customarily provided on a P.01 delay basis and due to the size of the aggregate loads mentioned above such groups varied in size from about 20 trunks upward. Any traffic overflowing from direct trunks to an alternate route via tandem would, therefore, require the addition of trunks to these rather large groups already established for the handling of traffic for which the tandem route was the first and only route. It now becomes pertinent to inquire into the capacity of trunks added to the tandem groups to carry rerouted load arriving from the direct high usage groups. From the P.01 trunk capacity table, it can be shown that adding a trunk to a group of 20 trunks increases the capacity of the group by 27 CCS, to the nearest whole CCS. Similarly adding one trunk to a 40 trunk group increases the capacity of the group by 29 CCS. It was decided in the New York City trials to use a constant average value of 28 CCS as the capacity of any trunks added to groups in an alternate route via tandem in order to accommodate rerouted traffic and still keep the groups on a P.01 basis for all traffic.

We are now in a position to compare the efficiency of a trunk added to a final route (two legs constitute one route) with the efficiency of a trunk added to the direct high usage group. The efficiency of the former is, for practical purposes, a constant, whereas the efficiency of the latter

is a function of the number of trunks provided to handle a given load on a lost-calls-cleared basis.

Referring once more to Fig. 2, let us assume that the ratio of the cost of an incremental path in ATN to the cost of an incremental path in AN is 1.4. It may then be stated that it costs 1.4 times as much to handle traffic on the alternate route as on the direct route. The cost ratio is computed on the basis of values of incremental trunks because the ultimate question in the economical separation of load between a direct and an alternate route is whether one more trunk should be added to the direct route or should more trunk capacity be provided in the alternate route to handle a marginal portion of the total load. Let us assume further that the offered load, A to N, is 240 CCS and that the efficiency of incremental trunks in the alternate route is 28 CCS per trunk.

With these three factors, the offered load, the cost ratio and the efficiency of incremental trunks in the alternate route, the most economical arrangement of trunks for carrying traffic from A to N may now be determined. The first step is to make sure that any trunk in the direct (HU) group will carry load at a cost per CCS equal to or less than the cost per CCS which is characteristic of the incremental trunks in the alternate route. Since the last trunk in a high usage group carries the least traffic, as previously discussed, the significant comparison is the ratio of the load carried by a trunk added to the alternate route to the load carried by the last trunk in the high usage group. The numerator of that ratio is 28 (CCS) and the denominator could be any one of 14 values shown on Curve A of Fig. 3, depending upon the number of trunks provided. If that ratio is made equal to the cost ratio (ATN/AN) there will be determined a value of load to be carried by a last trunk which in turn will determine the most economical number of trunks for the direct high usage group. This value is referred to as the "economic CCS" of the problem and is determined as follows:

$$\text{Cost ratio } \frac{1.4}{1.0} = \text{Efficiency ratio } \frac{28}{X}$$

$$X = 20, \text{ the economic CCS}$$

On Curve A it will be seen that the sixth trunk will carry 22.5, the seventh, 19.6 and the eighth, 16.4 CCS. Since the loading of the seventh trunk is closest to the economic CCS just computed, the conclusion is that seven trunks should be provided in the high usage group for the minimum overall cost of handling traffic from A to N. Since the seven trunks as a group will carry 185 CCS (Curve B), there will be 55 CCS

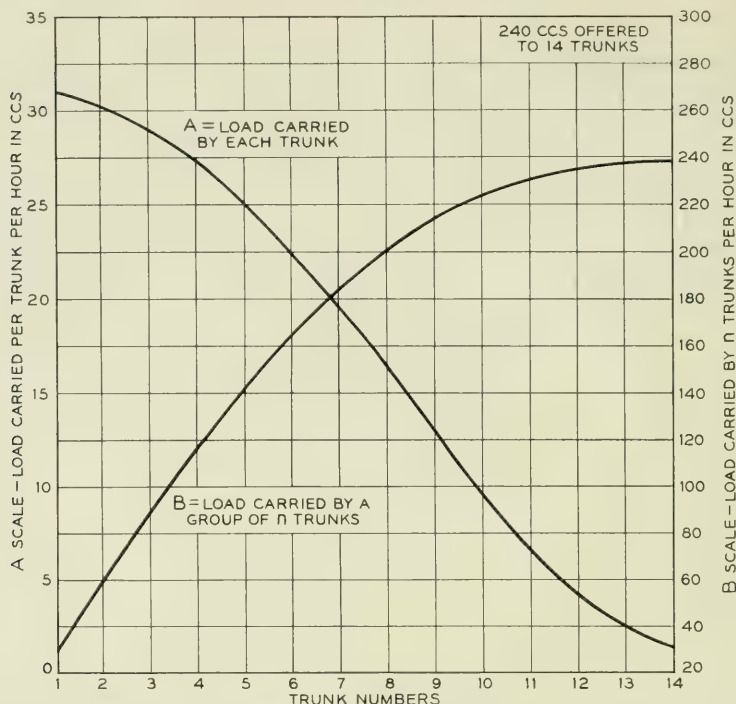


FIG. 3 — Distribution of offered load according to the Erlang B assumption, lost-calls-cleared.

(240-185) overflowed to the alternate route where they will be carried at the rate of 28 CCS per trunk. The trunks which need be added to the alternate route are therefore, $(55 \div 28)$ or 1.96.

The comparative costs of a number of alternate routing arrangements in which the only variable is the number of trunks provided in the direct high usage group are given in Table I. It is interesting to note that there is no measurable difference in cost between the optimum value for the number of HU trunks (7) and the cost of a 6 HU trunk arrangement. And indeed had five trunks or eight trunks been provided in the HU group under the stated conditions the additional cost above the most economical arrangement would have been nominal. This phenomenon is typical of many alternate routing triangles particularly when the offered load is in excess of say, 200 CCS. The significance of this fact is its effect upon the degree of accuracy with which cost ratios need to be determined. For example, assume that 1.4, the cost ratio used in our discussion, is precisely the right ratio for the facilities involved in the ATN triangle but that an approximate computation had arrived at a

value of 1.12. With the latter ratio the economic CCS would be $(28 \div 1.12)$ or 25. On Curve A of Fig. 3 it will be seen that trunk No. 5 (the last trunk of a 5-trunk group) carries 25 CCS. So, if the lower cost ratio had been used instead of the correct one, the effect upon overall trunking costs would have been 1 per cent in excess of the most economical arrangement. Had a cost ratio of 1.25 been used, six trunks would have been provided in the HU group and no cost penalty would have been incurred. On the other side of the optimum point an eight trunk HU group would meet the requirements of a cost ratio as high as 2.0 with a resulting cost penalty of about 2 per cent. As can be seen from Table I, the cost penalties mount more rapidly when more than the optimum number of high usage trunks are provided than when less are provided.

The principles of alternate routing and certain of the techniques used by traffic engineers in determining quantities and arrangements of inter-office trunks have just been described with particular reference to the trials that have been carried on in New York City. The latter were very extensive undertakings in which not only single alternate routes were provided but for a majority of items, multiple alternate routes. This was possible because New York City had two tandem systems each with a completing field to all city offices as well as other tandem systems (office selector tandems) each with a completing field to about 20 offices. Thus it was possible in many cases for an originating office to test a direct

TABLE I—COMPARATIVE COSTS OF ALTERNATE ROUTING SYSTEM
FOR VARIOUS ASSUMPTIONS AS TO NUMBER OF HU TRUNKS

Given: Offered load in CCS. 240
Efficiency of trunks added to alternate route. 28
Cost ratio, alternate to direct (HU) route. 1.4

No. of HU Trunks	Nominal Costs					% Deviation from Optimum
	HU trunks		Alternate trunks		All trunks	
	Per trunk	Total	Per trunk	Total added*		
3	1.0	3	1.4	7.50	10.50	7.70
4	1.0	4	1.4	6.10	10.10	3.59
5	1.0	5	1.4	4.85	9.85	1.03
6	1.0	6	1.4	3.75	9.75	0.00
7	1.0	7	1.4	2.75	9.75	0.00
8	1.0	8	1.4	1.95	9.95	2.05
9	1.0	9	1.4	1.30	10.30	5.64
10	1.0	10	1.4	.80	10.80	10.75

* Overflow CCS from HU Group $\times 1.4$

group, a group to an office selector tandem, a group to tandem No. 1 and finally a group to tandem No. 2 in succession in its quest for an idle path to a given office. It is not the purpose here to explore the intricacies of the multi-alternate route system but rather to point out that the basic principles described for separating a load between a first choice direct route and a single final alternate route on an economic basis were applied successfully to the multi-alternate route arrangements. A thorough analysis of one of the trials disclosed that there was a saving of approximately 20 per cent in the number of interoffice trunks of all types as compared with what would have been required without the use of alternate routes. Another result of the alternate routing system was an improvement in speed of service, since whatever portion of a load was carried by a high usage group was carried without delay and only the portion carried on the final alternate route was subject to about a 2 per cent delay. If, for example, 75 per cent of an item was carried without delay and the remainder at P.02 (two P.01 groups in tandem) the average delay for the whole load would be P.005 which is one half of the average per cent delay considered "normal" for a direct group without an alternate route.

ALTERNATE ROUTING IN INTERTOLL TRUNK NETWORKS

The successful application of alternate routing to local interoffice trunking led to investigation of the practicability of applying similar techniques to trunking between toll offices. The problem was different in a number of respects from that of the local interoffice system.

First, the physical size of the toll network covering the United States and Canada involved some 2,500 toll centers and lengths of haul between toll centers varying between 20 and 3,000 miles. From this it was evident that many first routes would not be direct but would involve one or more intermediate switching points thus complicating the problem of determining relative costs of first and alternate routes.

Second, the development of an alternate routing system for intertoll trunking would have to be predicated on the existence of a slow speed (high delay) trunking system. Whereas the aim of alternate routing within metropolitan areas was to make an already high speed trunking system cheaper the aim in toll would be to make a slow speed system a high speed system at little or no additional cost.

Third, new toll office switching facilities of a high degree of mechanical intelligence would be required.

Fourth, improvement in transmission and signaling design of toll circuits would be required to accommodate more links in tandem.

And fifth, an entirely new procedure for the correlation and routing

of more than two million items of traffic between toll offices, would need to be devised.

Over the years since 1945 practical solutions to the problems raised by the conditions have been attained by Bell System engineers and the fruits of their efforts will be put to the test in 1954 during which year the first practical application of "engineered" alternate routing in the intertoll network will be undertaken.

The remainder of this paper will be devoted to the more important aspects of the traffic engineering techniques used in determining the arrangement and numbers of intertoll trunks required in a multi-alternate routing system.

GENERAL TOLL SWITCHING PLAN

A brief description of the General Toll Switching Plan* will be appropriate here since any discussion of the alternate routing methods necessarily presumes an understanding of the basic pattern for routing traffic.

The plan under which transmission had been designed and traffic routings determined since 1930 comprehended a maximum connection of 5 intertoll trunks in tandem. Early studies of the alternate routing possibilities in toll networks led to the conclusion that a total of 8 intertoll

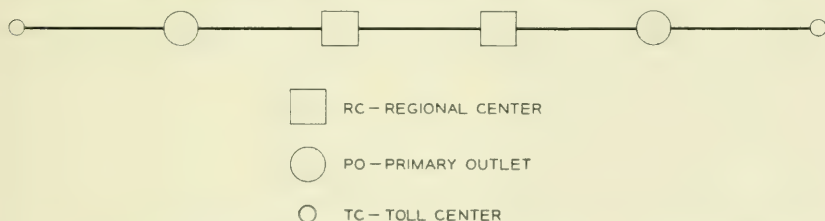


FIG. 4 — Illustration of present basic intertoll network showing maximum intertoll trunk linkage.

links would provide a more economical arrangement of trunks by shortening some very long final groups that would otherwise be required. Fig. 4 illustrates schematically the arrangement of switching centers in a maximum connection currently in use.

Fig. 5 shows a schematic of the proposed General Toll Switching Plan in which a connection involving 8 links is possible between TC1 and TC2. Such a route would constitute the final route between those TC's and each group in the route would be a low delay final group similar to

* J. J. Pilliod, Fundamental Plans for Toll Telephone Plant, Communications and Electronics, No. 2, Sept., 1952 and B. S. T. J., **31**, Sept., 1952.

the groups to and from tandem previously described (Fig. 2, AT and TN) for the local trunking situation. It should be noted here that, in the interest of economy, it was proposed for planning to further compromise with the lost-calls-cleared assumption by setting the speed objective on each final group at P.03 rather than P.01. Theoretical considerations indicate that the violence done to the Erlang B premise would not be severe enough to invalidate a study of trunk requirements made on that basis nor interfere substantially with the achievement of a high speed trunking system.

The switching plan requires that each switching office, i.e., the national center, the regional and sectional centers and the primary outlets be equipped with common control apparatus which would accept the number of digits (10) required by the nationwide numbering plan and direct a call to its designation in prescribed order over a variety of possible routes of which the high usage groups indicated on Fig. 5 are representative. (Alternate routes at TC's would be selected by operators.) The No. 4A toll switching system* was designed to perform this function and offices so equipped were designated control switching points or CSP's.

The CSP's were classified for transmission and traffic routing purposes in order of their relative importance in the network:

The national center (NC) has the unique function of switching traffic on the final route between regional centers thus being capable of handling traffic from any point in the country to any other point. In addition the NC is itself a regional center serving a particular area of the country;

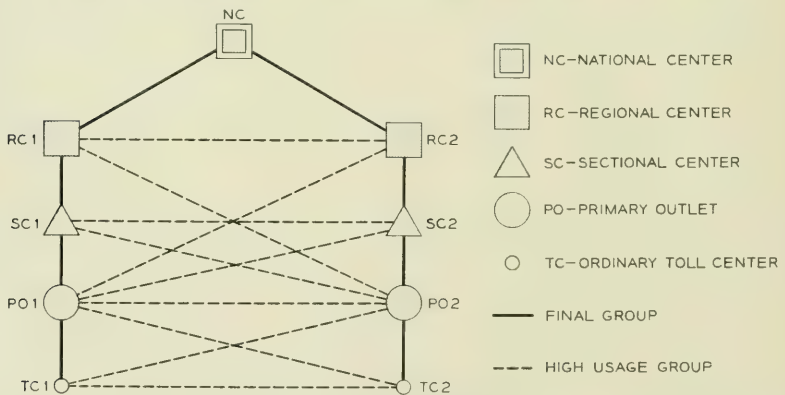


FIG. 5 — Illustration of basic intertoll network for nationwide toll dialing showing maximum final route linkage and typical high usage groups.

* F. F. Shipley, Automatic Toll Switching Systems, Communications and Electronics, No. 2, Sept., 1952.

Each regional center (RC) homes on, i.e., has a final group to the NC, and has one or more sectional centers homing upon it;

Each sectional center (SC) homes on an RC (or the NC) and has one or more primary outlets homing upon it;

Each primary outlet (PO) homes on an RC, NC or SC and has one or more ordinary toll centers homing upon it; and

Each ordinary toll center (TC) is so called because it performs no through switching function but merely serves as the connecting point between the intertoll network and local central offices or tributaries.

Thus each toll center (and in the generic sense this phrase includes all CSP's as well as ordinary toll centers) was classified with respect to the area served: the TC serving a group of local offices or tributaries, the PO serving a group of TC's, the SC serving a group of PO's, the RC serving a group of SC's and lastly the NC serving all the RC's. Under this arrangement any toll center could home on another of higher rank or classification. Thus a TC or PO, for example, could home upon an RC if so dictated by geographic and economic considerations. Before proceeding with a detailed study of trunk requirements the classification of toll centers and the homing relationships had first to be established. This was done in a manner which reflected the known densities and flow of traffic between the larger cities and the relative cost of final routes which in turn reflected the differences in lengths of haul to one CSP as opposed to another, etc. With the classification and homing of each toll center established it was possible to trace the final route, the route of last resort, between any two toll centers in the entire system. Thus was the stage set for determining the location of and number of trunks to be provided in high usage groups whose function would be to move traffic more economically by direct connection between points than could be done by following the final route.

It is apparent at once from the illustration in Fig. 5 that the problem of determining the most economical alternate for a given HU group is different from that encountered in the interlocal situation of Fig. 2 inasmuch as the latter had only one intermediate switching point in each final route. A further difference not specifically indicated is that intertoll trunk groups handle traffic in both directions whereas interlocal trunk groups handle traffic in only one direction, i.e., there are separate outward and inward groups between any pair of local offices. There are special circumstances under which one-way intertoll groups also are established but these may be ignored for purposes of our discussion.

While this paper is not specifically concerned with the transmission aspects of an alternate routing network some mention should be made

of that subject since it has an important bearing upon how traffic is routed. It is apparent that transmission standards for trunks connecting each class of office with any other class had to be so determined that a conversation between any two telephones in the system could be carried on within a specified limit of over-all transmission loss. For example, the possibility of a call being completed over an 8-link final route dictated the transmission tolerances to which each of the 8 links should be designed. Similarly, transmission equivalents for the various HU trunks over which calls would by-pass the final route had also to be determined. It became necessary, therefore, to devise a routing pattern which would assure that neither too many links nor the wrong classes of links would be connected in tandem. This interdependence of the transmission and routing aspects led to the adoption of the rule that traffic should be so routed that "when a call fails to find an idle path in the high usage network at any switching point (CSP) it can always be offered to the final network at that point." In other words, any CSP used as the intermediate switching point in the alternate route of a high usage group must lie in the final route between the terminal offices of the high usage group. This was the corner-stone of the routing pattern.

The Location of High Usage Groups

Having established the classification and homing arrangements of all toll centers and the basic routing pattern the next step in engineering the nationwide intertoll network was to determine between what points there should be high usage groups. It should be mentioned here that the following will describe in essential detail the procedures actually used by the Bell System in preparing the first coordinated nationwide estimate of intertoll trunk requirements under a plan deliberately designed to take advantage of the speed and economy possibilities of the alternate routing principles already described. The study, which was completed in 1951, involved a look into the future of about 10 years and was predicated on operator dialing of toll calls. Customer toll dialing which is foremost at the present time in Bell System planning may require some further modification of the general toll switching plan but the procedures about to be described will be substantially unchanged.

A high usage group is established between any two toll centers when the traffic between them in both directions is sufficient in volume to make such a group more economical than any other route available within the routing pattern. In the earlier discussion of the principles involved in determining what portion of a given load should be carried by a direct route it was shown that the ratio of the efficiencies and the ratio of the

costs of the direct and alternate routes are controlling. Since load carried by a high usage group is a function of the load offered, under the Erlang B assumption, it is apparent that the size of the load between any two toll centers will also limit the possible range of economic CCS values which can be realized. In other words a high usage group to exist at all must have at least one trunk and that one trunk must carry not less than the economic number of CCS required by the cost ratio applying to the case. For example, if a busy-hour load of 60 CCS is to be carried between offices A and B and the cost ratio indicates an economic CCS of 25 it can be shown that when one trunk is offered 60 CCS it will carry only 23 of the 60 CCS with the balance 37 CCS being overflowed. Under these conditions no direct (HU) group could be economically established since the efficiency of even the first trunk of such a group would fail to meet the requirement of the case.

This immediately suggests that a prime requirement in determining whether or not there should be a direct group of any size between two toll centers is a level of load at and above which a group will prove in and below which, of course, it will fail to prove in. As previously stated the cost ratio between the alternate route and the potential direct group is also a controlling factor. Thus, to determine the economic propriety of establishing a direct (HU) group it is necessary to know the following:

- Cost of path in the alternate route;
- Cost of path in the direct route;
- Efficiency of trunks added to the alternate route; and
- Load offered between toll centers.

The last three of these items were available to the engineer but the first item could not be known in all cases because the groups comprising the logical alternate routes in some cases were themselves hypothetical. For example, a high usage group between TC1 and TC2 in Fig. 5 might have an alternate route via P01 or via P02 which routes in turn would depend upon the existence of groups P01-TC2 and TC1-P02, respectively. Therefore, it was necessary to "cut-and-try" in the process of locating high usage groups. This was accomplished by choosing an average cost ratio (and hence an average economic CCS value) which could be used with the known offered loads between toll centers to test the feasibility of at least one high usage trunk. With this tentative pattern of high usage groups the potentially available alternate routes could then be identified. For each high usage-alternate route triangle thus tentatively selected the test of relative costs was applied to verify the economy of the case. Some proposed high usage groups failed to prove in under such test in which cases the uneconomic high usage groups were

abandoned and the loads that would have been offered to such groups were assigned switched routes. A switched route for our purposes is one in which the load between toll centers is carried on two or more inter-toll groups in tandem and may serve simultaneously as a first route for some items of traffic and as an alternate route for other items.

The cut-and-try process just described resulted finally in the location of all economically sound high usage groups. There remained, then, the problem of determining the number of trunks in each of those groups and in the final groups as well. At this point two rules were established affecting (1) the order of group load computation and (2) the direction of switched load concentration, respectively.

The first of these may best be described by considering the conditions illustrated in Fig. 6 where a portion of the intertoll network with typical homing arrangements and high usage groups is shown. The following assumptions regarding alternate routes may be made:

Traffic Item	First Route	Alternate Route
TC1-TC2	Direct	Via PO2
TC1-PO2	Direct	Via PO1
TC1-SC	Direct	Via PO1
PO1-PO2	Direct	Via SC

TC1 and TC2 are not switching centers and the load offered to their interconnecting group will be simply terminal traffic between their respective toll areas. Since TC1-PO2-TC2 has been determined as the most economical alternate route, calls from TC1 to TC2 overflowing the high usage group are offered to the TC1-PO2 group. Thus a part of the complete offering to TC1-PO2 is traffic rerouted from TC1-TC2. It follows then that the TC1-PO2 group can not be engineered until the overflow from the TC1-TC2 group has been determined. Similarly, the

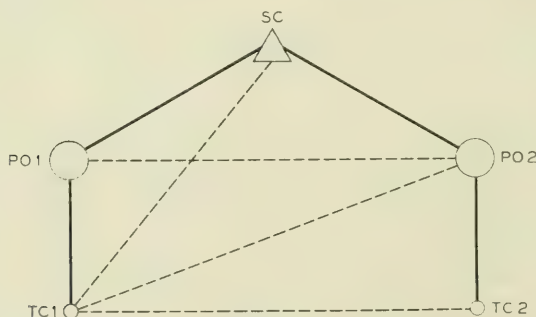


FIG. 6.

load offered to the PO1-PO2 group is composed in part of the overflow traffic from TC1-PO2. It is likewise clear that the final group TC1-PO1 will be offered overflow traffic from both TC1-TC2, TC1-PO2 and TC1-SC and hence the final group can not be engineered until the overflows from all high usage groups terminating at TC1 have been determined. This transfer of load from group to group with each succeeding high usage group serving as the base of a new triangle in orderly procession from TC to RC is the essence of the multi-alternate routing system and, therefore, a precise order of group load computation was necessary to assure proper accounting of the load offered to each group. In practice the order of group load computation requires that all TC-TC high usage groups be established first, then TC-PO groups, TC-SC and so on. The process involves selecting a particular TC as a "reference" and examining all traffic involving that TC to determine what (if any) high usage groups should terminate there.

The second rule, affecting the concentration of switched traffic, is related to the order of group load computation in that it postulates a starting point for the process of examining HU group possibilities. The need for such a rule may be best explained by reference to Fig. 7 representing a simple intertoll network in which PO, a CSP serving four TC's, homes on SC, another CSP, serving four other TC's. The question, answered by the second rule, is which of the eight TC's should be used as the first reference TC. The choice will determine the sizes of, i.e., the number of trunk terminations to be provided at, PO and SC, respectively. Let us examine the reason for this. Assume the TC's homing on PO are to be used as the first set of reference TC's (it makes no difference which TC is first chosen). Assume that the investigation showed an economical HU group between TC2 and SC but that no HU groups proved into any of the TC's (5, 6, 7, and 8) dependent upon SC. The load offered

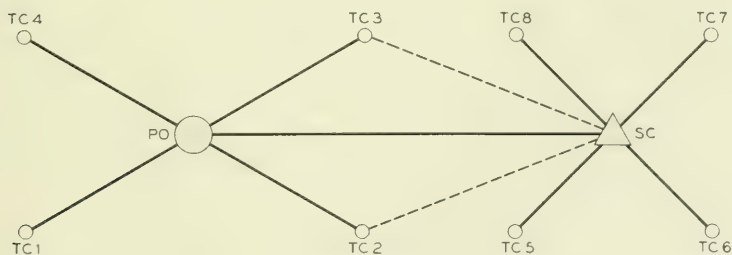


FIG. 7.

to the TC2-SC group, then, would consist of the sum of the following items:

TC2-TC5, TC2-TC6, TC2-TC7, TC2-TC8, and TC2-SC

Later examination of HU group possibilities with TC5 as the reference TC would have to recognize that the TC2-TC5 item was already committed to a switched route via SC. Thus the TC5 summation would consist of but four items (TC5-TC1, TC5-TC3, TC5-TC4, and TC5-PO) instead of five, thereby reducing the possibility of establishing a group between TC5 and PO. Even the four-item summation might prove in a group but if it be assumed further that a group proved in between TC3 and SC in addition to the TC2-SC group, the subsequent TC5 summation to the PO area would be reduced to 3 items, namely, TC5-TC1, TC5-TC4 and TC5-PO. It is clear that by using the TC's homing on PO as reference points the tendency is to build up the HU groups to SC and diminish the possibility of establishing HU groups terminating at PO with the result that SC handles more and PO less switched traffic. By thus selecting TC's homing on the lower class CSP, the PO, as the first set of reference points, an effect of "putting all eggs in one basket" is created since SC is already a larger switching office by virtue of its classification in the switching pattern. Therefore, it was deemed advisable to reverse the point of view in order to distribute the switched load more equitably among the various CSP's. To accomplish this each CSP was given an index number which reflected roughly both its size (total originating toll traffic) and its importance (classification) in the switching pattern. The most important CSP thus determined was assigned No. 1, the next most important No. 2, and so on throughout the list. Thus by starting with the TC's homing on the CSP with Index 1 as reference points, then moving to the TC's on Index 2 and so on the desired distribution of switching facilities was achieved.

GROUP BUSY HOUR VERSUS OFFICE BUSY HOUR TRAFFIC

Thus far we have considered the theory of alternate routing and the more important techniques used in applying it to the design of intertoll trunk networks. Another very important aspect of the whole problem involves a difference in load levels between any two toll centers during different hours of the day.

It has been the aim in engineering intertoll trunk groups without alternate routing to provide a sufficient number of trunks in each group to handle its traffic at a stated speed of service in the average busy hour of the busy season of the particular group. All significant traffic data for

administration and engineering have been obtained with that end in view. However, in alternate-routing networks the significant level of a trunk group load is not that which is characteristic of its own busy hour but rather that which is characteristic of the hour in which all trunk groups in a given network are collectively carrying the greatest aggregate traffic volume. For each group the former level is referred to as the group busy-hour (GBH) load and the latter, for convenience, as the office busy-hour (OBH) load.

The reason for using office busy-hour loads between toll centers rather than group busy-hour may be explained with reference to Fig. 8. Here is a represented part of an intertoll trunk network showing 4 HU groups connecting TC with four other toll offices a, b, c and d. TC homes on SC and for simplicity it may be assumed that the alternate route of each such group is the final group to SC.

In the hour during which the greatest volume of traffic is leaving and entering TC, i.e., the office busy-hour, the demand for trunk capacity in the TC-SC group will also be greatest since by design the final group to the home CSP of any TC is the route of last resort for all traffic to and from the TC. Thus the group busy-hour of the TC-SC group coincides with the busy-hour of TC as a whole.

The group busy hours of the respective HU groups (TC-a, TC-b, etc.) may occur outside of the office busy-hour for TC and during such hours the amount of traffic offered to and hence overflowed by each of the HU groups is greater on the average than that occurring in the office busy-hour. But at any other hour than the office busy-hour there is less total load on the network and hence there is some spare capacity in the final group available for handling the group busy-hour overflow of one or more of the high usage groups. By properly evaluating the average ratio between any given toll center-toll center load in its group busy-hour and its value in the office busy-hour it would be practicable to start with basic data in group busy-hour terms and convert it to equivalent office busy-hour levels before undertaking the procedures for separating loads between direct (HU) and alternate routes.

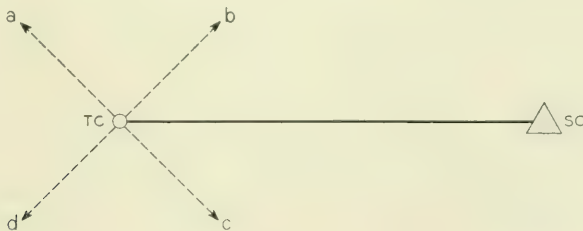


FIG. 8.

It became necessary, therefore, to develop these relationships from empirical data. Accordingly extensive field studies were conducted in a number of toll offices of different sizes to determine whether there was any consistent relationship between group and office busy-hour load levels of the many groups. Examination of group and office busy-hour loads on a total of almost 300 groups in 6 different offices showed a definite relationship which could be expressed as a function of a given value of group busy-hour load. Fig. 9 shows the group to office ratios finally computed from these data for use in the nationwide trunk study. The ratios for the smaller loads (30-90 CCS) showed considerable lack of consistency as between offices, but an adequate set of ratios for these

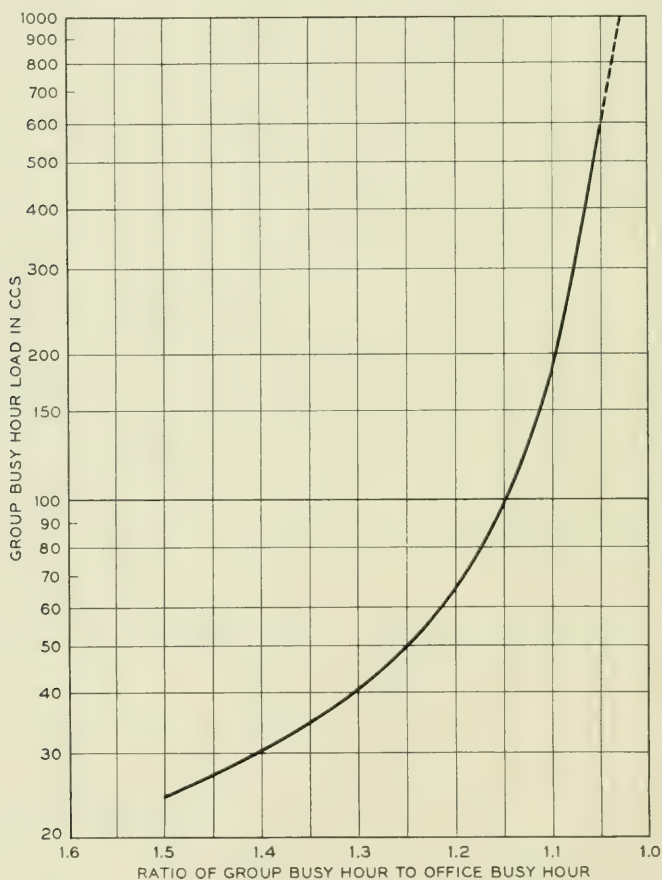


FIG. 9 — Relationship between group busy hour and office busy hour loads on intertoll trunk groups.

loads was arrived at through an averaging process. The ratio values are conservative, i.e., there was some evidence that in large toll offices the higher ratio values apply to larger loads than indicated by the curve. However, no attempt was made to develop separate sets of ratios for offices of different size since the use of the common set of values would tend merely to increase slightly the number of IU trunks required in a relatively few groups. The elimination of this small distortion did not appear to warrant the effort required to achieve it. The study data also showed that for a given toll office the sum of the respective group busy-hour loads for all groups was approximately 10-14 per cent greater than the sum of their respective loads in the office busy-hour.

The significance of this difference in the GBH and OBH aggregates is at once apparent. The intertoll trunk network designed on the alternate routing principle is required to handle from 10-14 per cent less busy-hour traffic than is required under the arrangement in which each toll center-toll center load stands alone and must be trunked for its own busy-hour. It is the pooling of trunk group capacities during the hour of maximum traffic flow for a given office plus the increased efficiency due to larger size of final groups that result in a requirement of fewer trunks in the network as a whole than would be required with any non-alternate routing trunking system of comparable service characteristics. The alternate routing system will enable the handling of normal busy-hour traffic on virtually a no-delay basis in so far as trunk provision may be controlling. The matter of speed of service potentialities and relative costs of alternate routing versus non-alternate routing systems will be treated later.

SOME UNANSWERED QUESTIONS

There are three questions upon the answers to which depend ultimate judgment of the adequacy and economy of a nationwide toll dialing network constructed upon the principles and with the techniques already described. These are:

1. What will be the effect upon trunk requirements of a proper evaluation of the non-random characteristic of lost calls, i.e., of the calls overflowed from high usage groups to other high usage or to final groups?
2. What will be the effect upon trunk requirements of a proper evaluation of the effect of non-coincidence of busy hours and busy seasons among the various toll centers?
3. What are the relative net costs of a nationwide intertoll dialing system engineered for multi-alternate routing and one designed without engineered alternate routing?

No completely satisfactory answers to these questions have yet been found but some knowledge of the answers can be gleaned both from theory and from much experience in the operation of intertoll trunk networks. Let us consider each question in turn.

Random Versus Non-random Traffic

Statistical theory states that the portion of random traffic which fails to be carried when offered to a given number of trunks is not itself random in nature. Such portion is, of course, the overflowed or "lost" calls described in the discussion of the Erlang B formula used in the separation of load between direct and alternate routes. Each such portion should be increased in volume to equal an equivalent random value before being offered to another trunk group since both the Erlang B and Poisson expressions of trunk capacity are predicated on the offering of random traffic. Such adjustment of high usage group overflows was not made for two reasons: one was the absence of any suitable increase factors for the variety of conditions encountered and the other was a definite indication from the New York City trials of alternate routing previously mentioned that such factors were not needed. Analysis of the New York trial indicated that when many parcels of overflow traffic are presented during one hour to the final route the aggregate tends to become random in nature. However, it was deemed advisable to provide some extra trunks in the final route to protect service by absorbing any abnormal peakiness in the offered traffic which might be due to non-random characteristics. As a practical matter it was possible to observe and regulate the loading of the final tandem route to the end that a satisfactory overall grade of service was maintained and from this and other subsequent experience to devise a rule-of-thumb for engineering future requirements in such final routes.

A similar approach to the problem was adopted in engineering the nationwide intertoll trunk network. An arbitrary increase was applied to the computed load offered to each final group. Whether this correction was too little or too much is still a moot question but it is believed that it was proper in the sense that it was in the right direction.

Statistical theory, however, does not support the view that a combination of non-random parcels tends to produce a random aggregate. The problem is under further study by the Bell Telephone Laboratories with a view to producing some practical means by which field engineers can readily and with reasonable accuracy adjust non-random trunk overflow traffic to equivalent random values for projection purposes.

Non-coincidence of Busy Hours and Busy Seasons

With respect to the second question involving the non-coincidence of busy hours among toll centers it should be noted first that trunk requirements estimated in the nationwide alternate routing trunk study were predicated on a common busy season and a common office busy hour for all toll offices and all intertoll groups. This premise resulted in system-wide requirements which were patently incorrect since it is known that different toll centers have different busy hours and that the busy season for toll traffic in New England for example is in the summer while that for Florida is in the winter. While it was possible to identify the busy season and the average busy hour of each toll center the statistical problem of incorporating such information for some 2,500 toll centers in a completely integrated nationwide study appeared insuperable.

The seriousness of the error introduced by the above premise is not as great as might at first appear since the New England Company should know the requirements for the busy season of its territory and the Southern Bell Company is equally interested in the busy season requirements for Florida. It is only in the trunks required for handling traffic between these two areas that a distortion of requirements could be readily demonstrated as a result of assuming the two busy seasons to be coincidental when in fact they are months apart. The example cited is an extreme case which serves to point up the problem. While no evaluation has been made of this distortion, and none seems statistically practicable, it is evident that the direction of distortion is toward over-estimation of trunk requirements.*

Thus it may be confidently stated that the general effect of assuming premise regarding the coincidence of busy seasons and busy hours upon the network as a whole was to compute trunk requirements in some groups more liberally than a precise evaluation of all significant factors would indicate as adequate. Proper evaluation of the effect of non-coincidence of busy seasons and busy hours will likely await the findings of field experience.

Costs — Alternate Routing Versus No Alternate Routing

In planning extensive and radical changes in the methods of handling toll traffic on a nationwide basis it was necessary to explore the economic

* In the New York City studies previously discussed, a similar assumption was made with respect to the coincidence of busy hours and busy seasons of the local offices. Due to the homogeneity of intra office traffic the degree of distortion in individual trunk group requirements was considered, except for a very few cases, to be insignificant.

soundness of the whole project. In the early stages of planning the bases for making such a judgment were rather meager since the only actual experience with alternate routing trunking had been in local systems, notably in New York City. Furthermore, the cost of the complicated switching mechanism, known to be essential to the plan but not yet developed, was an estimate of uncertain reliability. Added to these were the ever-present hazards of estimating the volume and distribution of toll traffic some 10 to 15 years in the future. However, in the light of the information then available it appeared that the overall cost of the proposed arrangements would be approximately equivalent to the cost of continuing present methods of engineering geared to a satisfactory trunk speed of service in the busy hour. Busy-hour trunk speed is defined as the average interval of delay* experienced in the busy hour by an attempt at trunk seizure. This interval excludes any effect of operating procedure as such and hence reflects only the adequacy of the trunk plant to handle the offered load. It should be noted parenthetically that the use of busy hour trunk speed as a criterion of adequate intertoll trunk provision is different from the criterion used for interlocal trunk provision. The former refers to the duration of delay encountered by the average attempt at trunk seizure whereas the latter concerns the percentage of attempts that will encounter delay of any duration. These different service criteria arose from the traditional need for high speed handling of large volumes of local traffic and the acceptability of a much slower speed for toll traffic.

The reference above to a "satisfactory" trunk speed of service presumes a stated objective to be achieved under the current method of engineering. In 1945 the service objective was expressed in different terms which can be interpreted as being roughly equivalent to a busy-hour trunk speed of 30 seconds. It is obvious that a nationwide network designed to produce an average busy-hour trunk speed, of, say 20 seconds would require more trunks than one designed to produce an average delay of 30 seconds. With an alternate routing system, however, there is only one speed of service under normal load conditions, namely, one in which the average delay is so small as to be incapable of meaningful expression in terms of seconds per average attempt. It follows then that in any attempt to assess the comparative costs of outside plant and equipment under the current method with those under the alternate routing method of engineering trunks it becomes necessary to first define the service level which the former is required to satisfy. Objective service

* S. C. Rappleye, A Study of the Delays Encountered by Toll Operators in Obtaining an Idle Trunk, B. S. T. J., **25**, Oct., 1946.

levels can and do change with the years for a variety of reasons thus automatically changing the cost comparison.

In spite of the difficulties in obtaining a true and stable comparison of the overall costs of the two methods the weight of evidence indicated the desirability of proceeding with the plans for achieving an ultimate goal of nationwide toll dialing employing the techniques of multi-alternate routing in the design of the intertoll trunk network.

With the completion of the study of the various intertoll trunk requirements for nationwide operator toll dialing there was established for the first time a bench-mark against which many of the assumptions, theories and procedures which went into its making could be measured for accuracy and practicability. Among these was the early question regarding the cost of operator toll dialing with engineered alternate routing compared to its cost without alternate routing. To arrive at a complete answer it would be necessary to restudy the entire network on the current basis of trunking, compare costs of dial switching equipment without CSP features with the cost of CSP switching equipment, evaluate changes in the location and types of trunk facilities and so on. The undertaking of a study and analysis of this scope would require a larger expenditure of engineering time and effort than would seem justified by the usefulness of the results. It should be noted, however, that analysis of the alternate routing trunk study indicated that the original premise as to trunk economies to be expected were substantially correct. In any event the advent of customer toll dialing with its peremptory requirement for high speed trunking has rendered the original question of the relative costs of operator toll dialing, with and without engineered alternate routing, somewhat academic. It can be safely assumed that a high speed intertoll trunking system suitable for customer dialing and engineered without alternate routing would be prohibitive in cost.

CONCLUSION

The transition from ringdown (wholly manual) handling of long haul toll traffic to operator dialing of such traffic has been proceeding for many years and at an increasingly greater rate during the last five years until now some 45 per cent of such traffic is dialed by operators. This has been accomplished almost exclusively on trunk networks operated without benefit of engineered alternate routing. Along with this, increasing use of dialing by destination code has been achieved as various cities and areas have converted to the nationwide numbering plan.

The second phase of the transition, now under way, involves the change from non-alternate routing to alternate routing trunking. With

this, as we have seen, will come a marked increase in the speed of movement of toll traffic.

The third and final phase of transition toward the ultimate goal now in its infancy, will involve progressive introduction of customer dialing of all toll traffic. Whereas, for operator toll dialing a high-speed trunking system is desirable, for customer toll dialing it is a firm requirement. Thus engineered alternate routing, which is the basis of the most economical and effective arrangements of toll plant facilities thus far conceived for highspeed trunking, will be indispensable to the further development of customer toll dialing.

There is little doubt that the unsolved problems of the moment will be resolved if not by statistical methods then by the empirical approach which has helped to find the answer to some telephone problems of the past. In the meantime, those who have been intimately associated with the project of nationwide toll dialing through these pioneering years have faith that the present plans and methods, revised as circumstances and experience may indicate, will result in the ultimate achievement of a toll service which will be fast, dependable and economical.

Intertoll Trunk Concentrating Equipment

By D. F. JOHNSTON

(Manuscript received August 25, 1953)

This article describes the Intertoll Trunk Concentrating Equipment which is a special purpose common-control type switching system. Its function is to combine, at a central point, small groups of trunks serving traffic originating at various outward toll switchboards and to route the combined traffic to a toll or tandem crossbar office in a distant toll center. The operators at the outward toll switchboards are thereby provided with the equivalent of direct access to the intertoll trunk circuits.

Both operating and equipment savings are realized by the use of this concentrating equipment as compared to handling the same traffic via a toll crossbar office at the originating toll center.

INTRODUCTION

This article deals with a method of handling traffic from outward toll switchboards in a metropolitan toll center to a specific distant toll center with the objectives of (1) providing the equivalent of direct access to intertoll circuits from the individual switchboards, for traffic for which direct circuits cannot be justified, (2) giving relief to the No. 4 type toll switching system in the metropolitan toll center, and (3) providing a means for the dispersion of toll switching facilities.

A metropolitan toll center may contain a number of outward toll switchboards. Some of these are situated in the central toll building and others in decentralized locations. An individual outward switchboard may have a sufficient amount of traffic to a specific distant toll center to justify a group of intertoll circuits direct from the switchboard to the No. 4 type toll or crossbar tandem office in the distant center. It has been the practice to provide such direct access to intertoll circuits at centralized toll switchboards. Traffic exceeding the capacity of such intertoll trunk groups is handled over tandem trunks from the toll switchboards via the No. 4 type toll crossbar office in the originating center. The decentralized switchboards in general have reached intertoll cir-

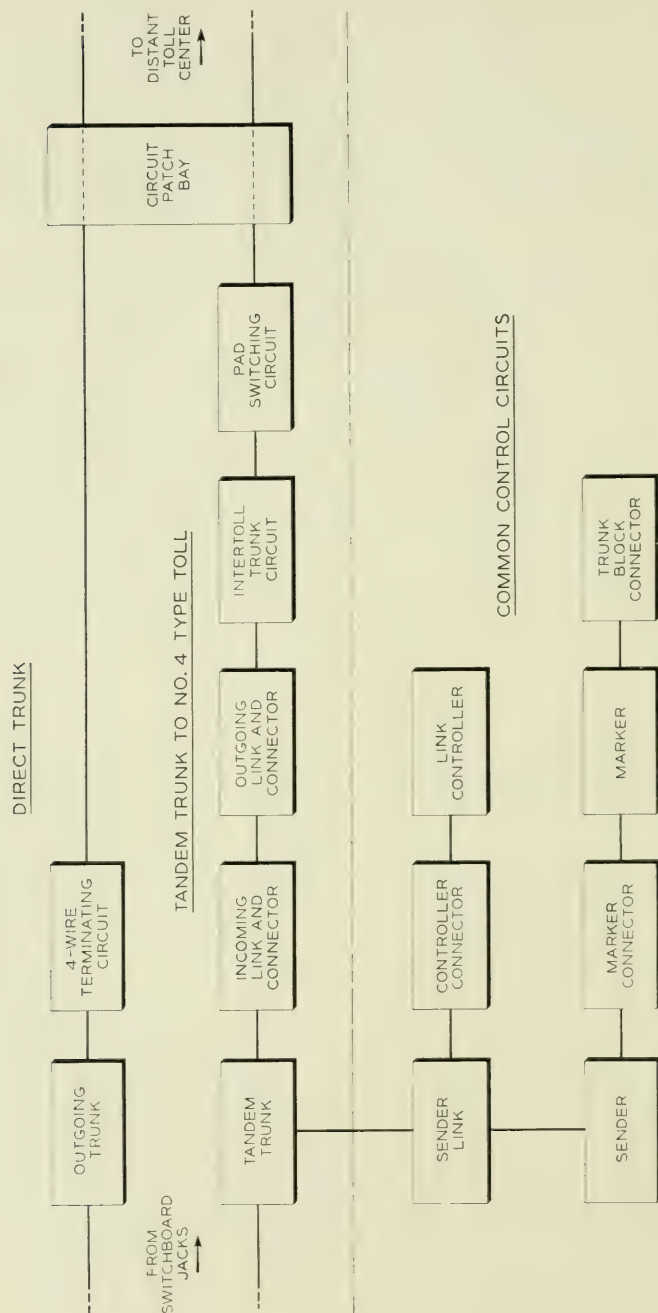


Fig. 1 — Switching facilities used in establishing connection to intertoll trunks — direct access versus access via toll crossbar office.

cuts only via tandem trunks to the No. 4 type toll crossbar office in the home toll center.

Handling such traffic through the toll crossbar office involves greater operating effort than handling it by direct trunks since three additional digits per call must be keyed by the operator to direct the call through the toll crossbar office. Also the cost for an intertoll connection is higher with the tandem trunk method than with the direct circuit method, due to the greater number of switching facilities used in establishing the connection. Fig. 1 shows the components used in both cases up to the point where they join common facilities.

INTERTOLL TRUNK CONCENTRATING EQUIPMENT

General

The intertoll trunk concentrating equipment has been developed to provide the equivalent of direct access to intertoll trunks for the traffic, from individual outward toll switchboards, which cannot justify the use of direct trunks. It is a small special purpose common control type switching system located at a central point. It gathers small traffic loads, to a specific destination and automatically routes this traffic to a common group of intertoll trunk circuits which terminate in a toll or tandem crossbar office in the distant toll center.

The maximum capacity of a trunk concentrating equipment is 100 incoming trunks and 40 outgoing trunks. It may be furnished in smaller sizes. If more than 40 outgoing trunks are required to a particular destination, additional concentrating equipments may be furnished.

The field of use for this equipment lies between that of direct trunks and trunks reached through the toll crossbar switching system.

The concentrating equipment is arranged only for multifrequency pulsing from the switchboard. This is a system of pulsing in which combinations of two frequencies within the voice frequency band are transmitted over the talking path to the distant end. Each digit from 0 to 9 employs a different pair of frequencies.

The intertoll trunk concentrating equipment consists of 4 basic circuit components, namely, incoming trunk, trunk selection switches, controller and outgoing trunk circuits.

The detached contact form of circuit presentation is employed in the figures because of its simplicity. In this method the core and winding of a relay may be shown in one location and the associated contacts in other convenient locations. The core and contacts are related by the common designation which appears at the symbols which represent them.

Incoming Trunk Circuits

Several types of incoming trunk circuits are provided to meet the various switchboard conditions where a switchboard and concentrating equipment are located in the same building or to meet the interoffice trunk loop conditions where the switchboard and concentrating equipment are in separate locations. In all cases the incoming trunk circuit receives signals from the switchboard, directly or indirectly and transmits them to the outgoing trunk circuit and to the controller. The incoming trunk circuit also receives signals from the outgoing trunk circuit and transmits them to the switchboard. The incoming trunk circuits, where necessary, also convert the two-wire transmission circuit from the switchboard to a four-wire transmission circuit to meet the intertoll trunk facilities. Fig. 2 shows a typical incoming trunk circuit.

Trunk Selection Switches

The trunk selection switches are the elements which, under the direction of the controller, connect the incoming trunks, requesting service, to the outgoing trunks. Crossbar switches having 20 verticals and 10 horizontals, providing 200 crosspoints, are used. A crosspoint is the intersection of a vertical and a horizontal element of the switch. The incoming trunks are connected to the verticals and the outgoing trunks to the horizontals of the switch. By the switch operation any of the 20 incoming trunks may thus be connected to any of the 10 outgoing trunks.

A single switch will accommodate 20 incoming and 10 outgoing trunks. To increase the number of incoming trunks a similar switch is added for each additional 20 trunks with the verticals connected to the new incoming trunks and the horizontals connected to the corresponding horizontals of the first switch. To increase the number of outgoing trunks switches are added with the horizontals connected to the new group of outgoing trunks and the verticals connected to the corresponding verticals associated with other groups of incoming trunks. The switch arrangement may be envisioned as one large switch having as many verticals as there are incoming trunks and as many horizontals as there are outgoing trunks.

Outgoing Trunk Circuit

Only one type of outgoing trunk circuit is required, as shown in Fig. 3. All outgoing trunks from a particular concentrating equipment terminate in the same distant toll center. The outgoing trunk circuit receives

signals from the incoming trunk circuit to which it is connected and transmits them to the intertoll facilities and vice-versa.

Four-Wire Switching

Hybrid coils are provided in the concentrating equipment to (1) couple the two-wire transmission incoming trunk circuit to the four-wire transmission outgoing trunk circuit and (2) match the impedance of the incoming trunk to that of the outgoing trunk.

The impedance of each two-wire leg of the outgoing trunk is always 600 ohms. The impedance of an incoming trunk may be 600 ohms or 900 ohms or 1,500 ohms. Thus three hybrid coil ratios are required namely 1 to 1, 1.5 to 1 and 2.5 to 1.

Since the hybrid coil and the associated balancing network are determined by the particular incoming trunk with which they function they

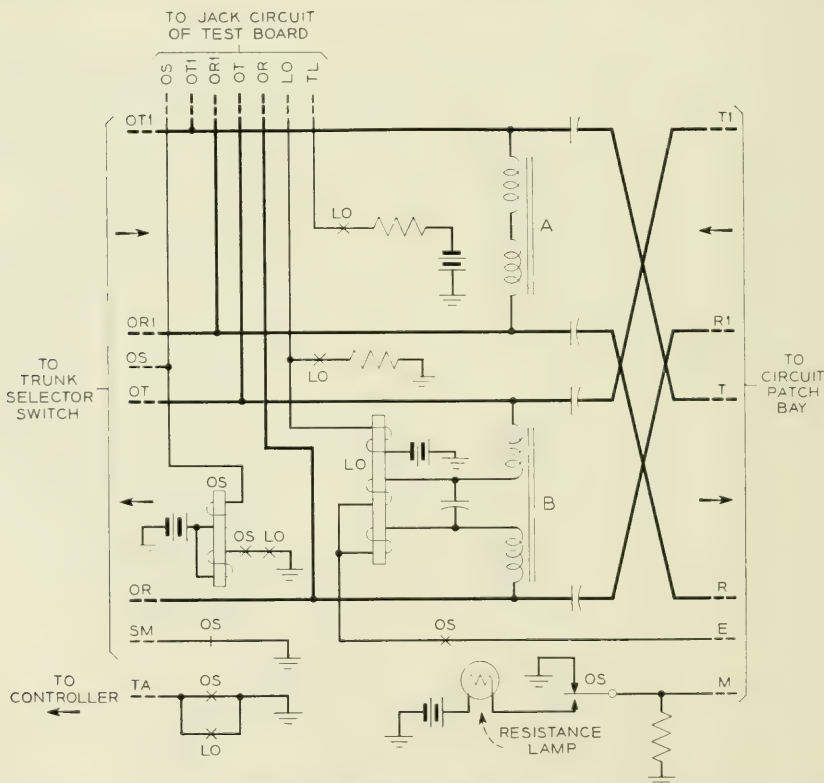


Fig. 3 — Outgoing trunk circuit.

are made a part of the incoming trunk circuits rather than a part of the less numerous outgoing trunks to which they may be switched. Switching between incoming and outgoing trunks is therefore performed on a four-wire basis.

Controller Circuit

The controller circuit is the control element of the trunk concentrating equipment. Its primary function is to select the incoming and outgoing trunks to be interconnected and to operate the proper select and hold magnets of the associated switches to close the required crosspoint. Since the controller circuit is described in some detail later, only the general principles of it will be dealt with at this time.

The controller divides the incoming trunks into ten groups, (0 to 9), of ten trunks each. When idle it admits calls for a very short interval and then closes gates which exclude all other groups until calls recognized, within the gate, have been served. The controller serves the groups and trunks within the gate in one of two orders depending upon the direction of selection existing at the time. In one case selection will start with the lowest numbered trunk in the lowest numbered group and progress to the highest numbered trunk in the highest numbered group. In the other case the order will be reversed starting with the highest numbered trunk in the highest numbered group and progressing to the lowest numbered trunk in the lowest numbered group.

The controller also divides the outgoing trunks into four groups, (0 to 3), of ten. The outgoing trunks are served in order, either from low numbered to high numbered trunks or vice-versa. Once selected, the outgoing trunk remains locked out, after use, until all trunks have been used, or until a trouble condition causes a reversal of the direction of selection of the outgoing trunks.

To avoid connecting an incoming trunk to two outgoing trunks or connecting two incoming trunks to an outgoing trunk the controller tests both the select and hold magnets for possible trouble conditions, such as crosses, before operating them.

Each intertoll trunk concentrating equipment has but one controller. If the controller ceased to function the entire concentrating equipment would be out of service. To insure reliability the philosophy was adopted in the design that no single trouble should disable the controller. This accounts for some of the features, the reasons for which otherwise are not obvious.

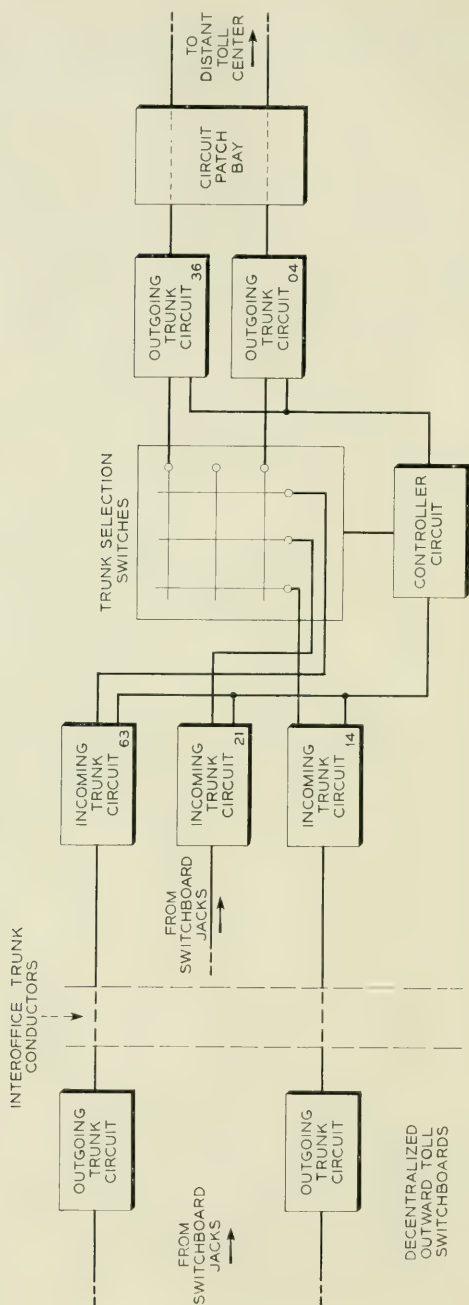


Fig. 4 — Circuit components of intertoll trunk concentrating equipment.

Completion of Calls

The basic circuit components of the intertoll trunk concentrating equipment are shown in Fig. 4. Three incoming trunks in different groups and from different switchboards are shown. Two outgoing trunks in different groups are also shown. Assume that the controller is conditioned to serve both incoming and outgoing trunks in the low to high order, that a call is placed on incoming trunk 14 and a short time later on incoming trunk 21, and that outgoing trunks 04 and 36 only are available for selection. Seizure of trunk 14 at the originating switchboard causes the incoming trunk to place a ground on the start lead of incoming trunk group 1. This ground indicates to the controller that a trunk in incoming group 1 is calling for service. The controller then closes the gate to all other ten groups, tests the select magnets associated with horizontal 04 for crosses (since outgoing trunk 04 is first in order for selection) and finding none operates the select magnets. It then tests the hold magnets associated with vertical 14 and finding no crosses operates the hold magnets, thus closing the crosspoint whose coordinates are (04, 14) connecting incoming trunk 14 to outgoing trunk 04. The outgoing trunk transmits a seizure signal over the intertoll facilities to the distant end which then transmits a signal back to incoming trunk 14, which relays the information to the controller. The controller then releases from that connection, opens the gates and admits the call waiting on incoming trunk 21. It proceeds to complete this call to outgoing trunk 36 which is next in order of selection in a similar manner.

If in the assumed case outgoing trunk 36 was the last trunk then available for selection the controller would, at the completion of selection of trunk 36, proceed as follows:

(1) If any of the other outgoing trunks which had been locked out were idle the controller would now make these trunks available for selection.

(2) If no trunks were idle the controller would wait, and cause a signal to be transmitted to all associated outward switchboards. This signal will prevent the lighting of idle trunk indicating lamps at each switchboard. When one or more outgoing trunks become idle the controller will make them available for selection and will permit the idle trunk indicating lamps at the associated switchboards to light as an indication that trunks are available.

Component Circuits of the Controller

In the following paragraphs the component circuits of the controller will be described individually. The descriptions of these circuits contain the minimum of detail required to understand how they function.

"Tens Group" and "Tens Gate" Relays for Incoming Trunks

The incoming trunk associated with each controller are divided into groups of ten. These groups are designated 0, 1, 2, etc. Each trunk in the group has a unit designation 0, 1, 2, etc. which corresponds to the vertical of the switch to which it is connected. For instance, an incoming trunk in group 2 which is connected to the number 3 vertical on the switch is designated 23. Each group of trunks has a common start lead, (sr), (See Fig. 2) to the controller which is grounded when any trunk in the group is calling for service. Each trunk in the group of ten supplies an individual lead *us* to the controller which is also grounded when the particular trunk is calling for service. This lead serves to identify the units designation of the trunk.

In the controller there is a group of three relays associated with each group of 10 incoming trunks (Fig. 5). These are (1) the TENS relay (TX-) which responds to the ground on the *sr* lead from the trunk group when a trunk in that group is calling for service providing that the TENS gate is open as discussed below, (2) the UNITS CONTROL relay (UC-) which when operated connects the common group of UNITS relays (U0-U9) to the *us* leads of the trunk group, (3) the HOLD CONNECT (HC-) which when operated steers the hold magnet operating path to the hold magnets associated with the trunks in the particular group. The UC- and HC-relays do not operate until it is the turn of that associated trunk group to be served.

Two series chains, carried through transfer contacts on all of the TENS relays, control the operation of the UNITS CONTROL and HOLD CONTROL relays. The operation of these relays and the manner of advancing selections from one group to another is best explained by the use of the following example. Assume that incoming trunks in groups 1 and 3 have originated calls resulting in the operation of the TX1 and TX3 relays. Assume also that the controller is conditioned for the low to high direction of selection for incoming trunks. When the TX1 and TX3 relays operate, they cause the release of the TENS GATE relays (TG1 and TG2) which close the gates to the operation of any other TX relays and operate the UC1 and HC1 relays. When the UNITS gates close as described later, the TX1 relay is released and the trunks in group 1 will be served. When the last trunk in this group has been served the UC1 and HC1 relays release and the UC3 and HC3 relays operate advancing the selections into group 3. The TX3 and UC3 and HC3 relays then function in the same manner as described for the TX1, UC1 and HC1 relays. If the direction of selection had been from high to low instead of from low to high, group 3 would have been served first instead of group 1.

Grouping and Individual Relays for Outgoing Trunks

In the controller the outgoing trunks are divided into groups of ten circuits. This is done to keep the length of chains on the control relays short (ten rather than forty relays long) and to reduce current drain during light periods since a possible maximum of nine relays will be locked up instead of 39. For each outgoing trunk one OUT TRUNK (OT-) relay (shown in Fig. 10) is provided in the controller and for each ten of these a GROUP relay (GP-) is provided (shown in Fig. 11). When the connection of an incoming trunk to an outgoing trunk is completed the controller operates the OT relay, corresponding to the outgoing trunk, which advances the selection path to the next idle trunk in the group and locks up under control of the GP relay for that group. When the last idle trunk in the group has been selected the group relay is released advancing the selection path to the next group and releasing all OT relays in its own group which are associated with idle trunks. These trunks are then locked out of service until all trunks above or below it (depending upon the direction of selection) have been selected. Two chains are carried through contacts of the group relays and through the OT relays of each group for controlling the operation of the switch select magnets, one for each direction of selection. These chains are shown in Fig. 7.

Units, Unit Gates and End Relays

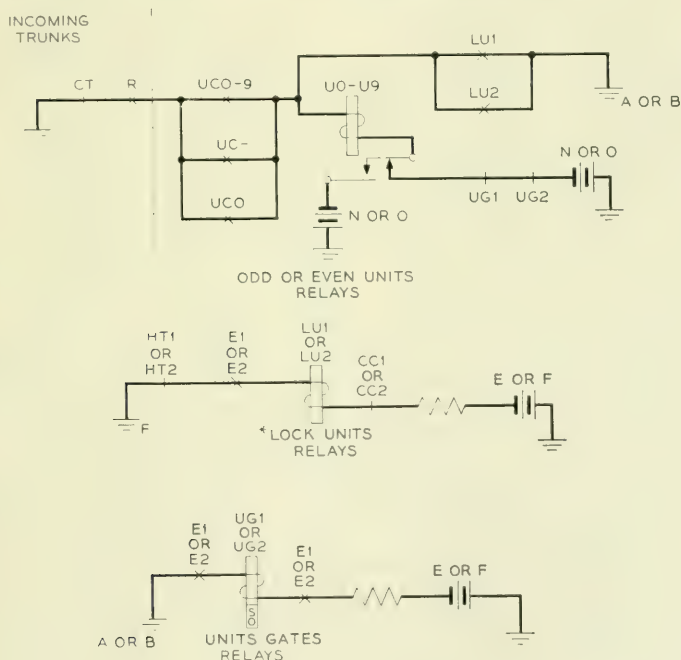
Ten Unit relays designated τ_0 to τ_9 are provided to identify the individual incoming trunks in a group of ten which have calls waiting to be served. These relays, together with the UNIT GATES and LOCK UNIT relays are shown in Fig. 6.

When the UNITS CONNECT relay (UC-) associated with a group of trunks operates as previously described, it connects these ten τ - relays to the u -leads from the incoming trunks. Those of the τ relays which find ground on these leads, indicating calls waiting to be served, operate and lock. Any u relay operated operates one of two END relays E1 or E2. The END relay in turn operates one of two UNIT GATE relays (UG_1 or UG_2) which opens the operating, but not the locking, path of all u relays thereby preventing a late arrival from stealing preference from an earlier originated call. One END and one UNITS GATE relay are provided for each direction of selection. The hold magnets associated with the incoming trunks in the group being served are connected through contacts on the HOLD CONTROL (HC-) relay to two chains of transfer contacts on the UNIT relays (1 chain for each direction of selection). As the connection of the incoming trunk is completed to the outgoing trunk the UNIT relay

associated with that incoming trunk is released advancing the selection path to the next incoming trunk to be served. When the last incoming trunk within the gate has been selected the END relay releases, signifying the end of selections in that group and enabling the controller to advance to the next group. The release of the END relay releases the UNITS GATE relay restoring the operating path for all U relays.

Select Magnet Operating Circuit

The select magnet operating circuit is shown in Fig. 7. Two of these circuits are provided in accordance with the philosophy that no single trouble should block the concentrator. One of these circuits is associated with each direction of selection. When the UNITS GATE relay operates as described in the preceding paragraph, battery is connected through the windings of the XS and SS relays, chains on the group relays (GP-), chains on the OT relays, to the select magnet associated with the lowest



* LOCK UNITS RELAY DESIGNATIONS:
 -- 1 ARE INCOMING LOW TO HIGH
 -- 2 HIGH TO LOW

Fig. 6 — Units, unit gates and lock unit relays.

or highest unoperated OT relay in the group depending on the direction of selection. If the path is found to be closed to ground the ss relay will operate. The xs relay is a polar relay. The resistance shown in series with the secondary winding will have a value depending upon the number of select magnets which are multiplied together on each level of the particular concentrator, which number is a function of the number of incoming trunks. The xs relay will operate on every normal connection since the current in the s winding, which is in the direction to operate the relay, will be larger than that in the primary winding. When the resistance to ground on the select magnet lead is less than it should be, due to a cross with another select magnet lead or to a direct ground, the current in the primary winding will be greater than that in the secondary winding and the resultant ampere turns will be sufficient in the non-operate direction to prevent the operation of the xs relay. This would cause the selection to be halted, the controller to time out and the trouble registered. A reversal in the direction of selection would occur and selections would then be resumed. If no fault is found with the select magnet operating path the select magnets operate. The controller then introduces a small time interval to permit all parts of the selecting bar operated by the select magnet to come to rest. The hold magnet operating path is then closed as discussed below and the crosspoint is closed. The closure of the crosspoint operates a relay os in the outgoing trunk which in turn releases the select magnets and ss relay which releases the xs relay.

Hold Magnet Operating Circuit

The hold magnet operating circuit is shown in Fig. 8. Two circuits are provided, one for each direction of selection to insure against blocking in case a trouble in this portion of the controller. The cross detection part of this circuit is an unbalanced wheatstone bridge, the galvanometer element of which is the polar relay xh. Three of the arms of the bridge are resistances, the values of which are tailored to each particular concentrator depending upon the number of hold magnets to be encountered on a normal connection. This number is a function of the number of outgoing trunks. The fourth arm consists of the hold magnets. When the TENS GATES relays released as previously described the xh relay operated (at this time the hold magnets are not connected and the bridge is not formed). Later in the progress of the call, when the UNITS relays operate, the hold magnets are connected and the bridge is formed. If the resistance of the hold magnet arm of the bridge is as expected, the bridge is unbalanced so as to keep current flowing through the xh relay in the direction to maintain it operated. This will permit a relay ur, which furnishes

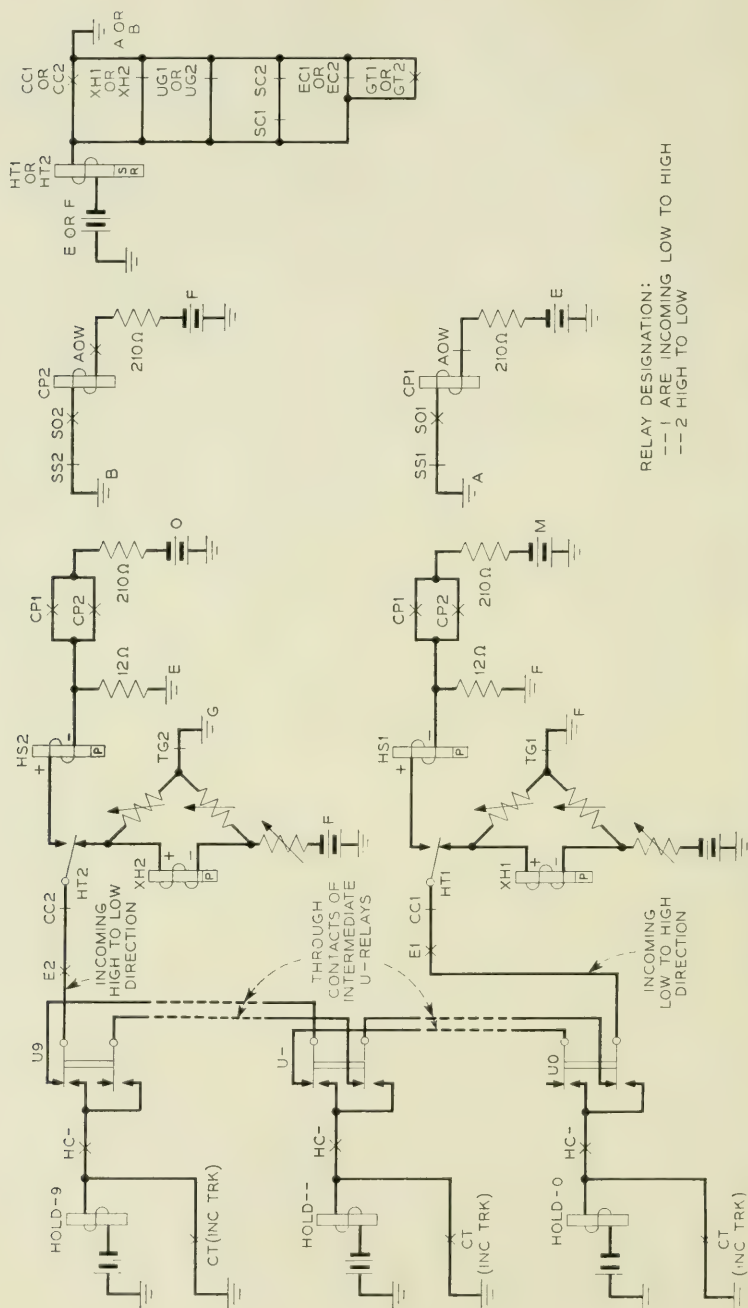


Fig. 8 — Hold magnet cross detecting and operating circuit.

the time interval, between the operation of the select and hold magnets, to release and the connection to proceed. If, however, the hold magnet lead is crossed to another hold magnet lead the resistance in the hold magnet arm of the bridge will be lower than expected causing the XH relay to release. The release of this relay will prevent the release of the timing relay, thus causing a time out and trouble registration.

If no trouble cross is found on the hold magnet lead the HT relay releases when the select magnet is operated. This is a slow releasing relay and provides the time for allowing the selecting bar to come to rest before the hold magnets are operated. When this relay has released low resistance ground through the HS relay and 12-ohm resistance is connected to the hold magnets as an operating path. The HS relay is polarized and the current flowing from the hold magnets is in the direction to non operate the relay. The operation of the HS relay will be discussed below.

Outgoing Trunk Seizure and Continuity Check

Up to this point the select and hold magnets have been operated. The crosspoint of the switch is now closed and the outgoing trunk seized.

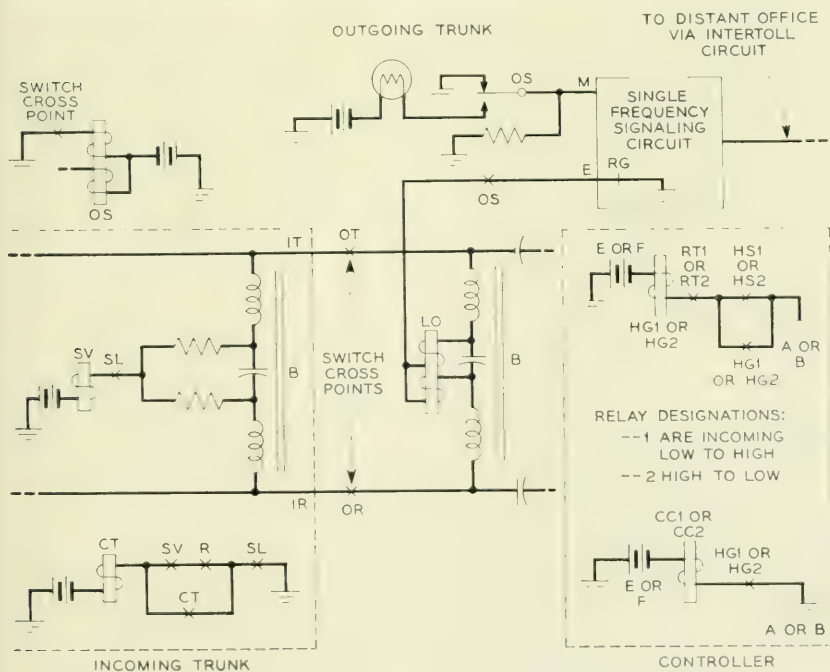


Fig. 9 — Outgoing trunk seizure and continuity check.

To insure that continuity exists through the switch a check is made by making the advance of the controller dependent upon receiving a signal from the distant end. The os relay, Fig. 9, in the outgoing trunk is operated directly by a ground on one of the contacts of the crosspoint, and opens the operating path of the select magnets and ss relay (shown in Fig. 7). The release of the ss relay causes the operation of the cp relay shown in Fig. 8. The operation of the cp relay changes the potential connected to the winding of the hs relay from ground to about -2.5 volts. The operation of the os relay is the outgoing trunk sends a seizure signal over the trunk to the distant end which then sends a stop signal back to the incoming trunk of the concentrator causing the operation of the SUPERVISORY RELAY (SV) and the CUT THROUGH relay (CT) of the incoming trunk which locks under control of the operator and connects ground to the hold magnets of the switch. This ground causes the current through the hs relay to be reversed, operating the relay, indicating a successful continuity check. A failure to make a successful check will cause the controller to time out, lock out the selected outgoing trunk and proceed to the next idle outgoing trunk.

Outgoing Trunk Lock Out and Advance

A TRUNK ADVANCE (TA) relay shown in Fig. 10 is provided for each group of outgoing trunks. The (TA) relays operate in multiple on every call. They are operated when the select magnets are energized and are released, under normal conditions, when a successful continuity check is made, or in the event of an unsuccessful check, when the controller times out. Their function is to prevent the operation of the ot relay associated with the outgoing trunk until a connection has been suc-

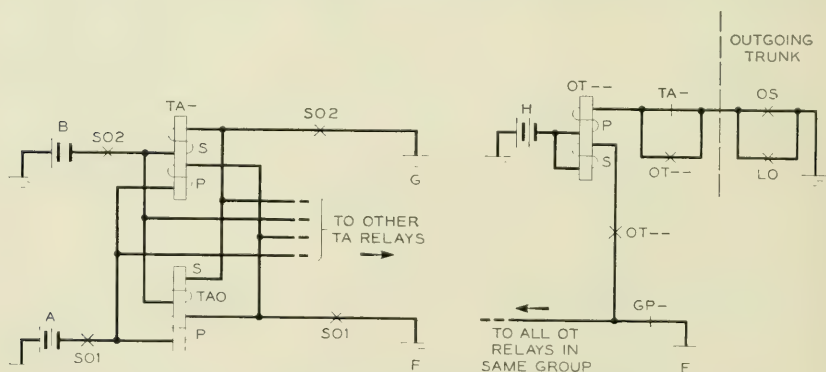


Fig. 10 — Outgoing trunk lockout and advance.

cessfully completed except as stated above where a continuity check has failed. When the TA relays release, the OR relay associated with the trunk selected, or passed by, is operated and advances the selection to the next idle trunk. The operated OR relay locks to its own GROUP relay GP until the last OR relay in that group has operated. When the last trunk in that group is selected the GP relay of that group is released, advancing the selection path into the next group. The OR relays in the first group are then controlled only from the individual outgoing trunk with which they are associated and will be released when the trunk becomes idle. The trunks in that group cannot be selected again until the group relay has been reoperated which can not occur until all of the outgoing trunks on the concentrator have been selected in turn or until the controller times out. In both of these cases the controller will go through END OF CYCLE operation and reoperate the group relays as later described. With the above arrangement the traffic is spread evenly over the whole group of outgoing trunks.

End of Cycle and Guard Timing

When a connection through the concentrating equipment is released by an operator, the outgoing trunk sends a disconnect signal to the other end, to release the equipment there. It may take approximately 0.75 seconds for the distant equipment to release after it has received the disconnect signal. If this equipment should be re seized before it is released the new call would be connected to the same subscriber. A guard time could be incorporated in every outgoing trunk to prevent the transmission of a seizure signal for this 0.75 seconds, but this would be relatively expensive. To avoid this procedure the end of cycle and guard timing feature has been incorporated in the controller. This feature, shown in Fig. 11, together with the outgoing trunk lock out feature insures that no outgoing trunk can be seized for at least 0.8 of a second after it has been released from a previous call.

When the last available idle outgoing trunk on the concentrator has been selected the last operated group relay GP releases, and two END OF CYCLE relays EC1, EC2 which are normally operated also release. Either one of these relays released operates both the GROUP RESTORE relay GR and an ALL BUSY relay AB. Both the GR and AB relays cause the idle indicating lamp associated with trunks to the concentrator at all originating switchboards to remain dark as an indication that all trunks are busy. If no trunk outgoing from the concentrator is idle at this time nothing else will happen in the controller. When any outgoing trunk in any group becomes idle the GROUP relays associated with such groups

will operate. Any GROUP relay operated will operate two GUARD TIME relays GT1 and GT2 which will reoperate the END OF CYCLE relays and enable two timing circuits. The END OF CYCLE relays operated release the GROUP RESTORE (GR) and ALL BUSY (AB) relays restoring control of the idle indicating lamps at all originating switchboards to the associated trunks. Timing is obtained by means of cold cathode gas tubes in combination with resistance-capacitance circuits. The control gap of the gas tube (terminals 1 and 4) is connected across the capacitor which is normally short circuited thru 510 ohms. To start timing, the GT-relay remove the short circuit around the capacitor and connects ground thru the TF-relay in series with the capacitor and the 0.556 megohm resistor. The capacitor charges from the 130-volt battery. When the voltage across the capacitor reaches the breakdown potential of the gas tube the latter fires and operates the TF relay. The time to fire the tube is determined from the following equation:

$$T = 2.3 RC \text{ Log } \frac{E}{E - V},$$

where E is the voltage of the battery, V is the voltage at which the

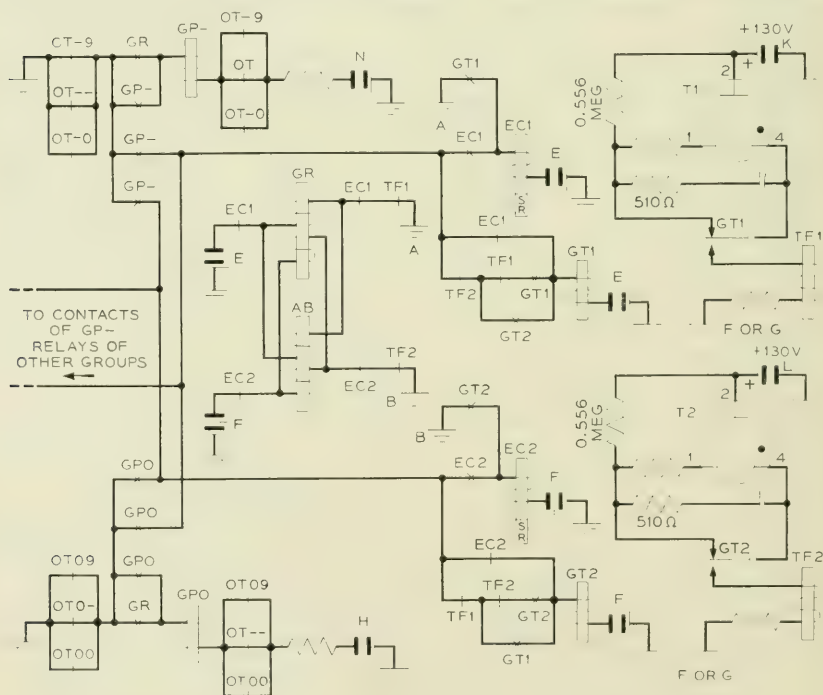


Fig. 11 — End of cycle and guard timing.

control gap breaks down (which is minimum 63 volts for this tube) R is the value of resistance in ohms, and C is the value of the capacitor in farads. The timing circuits measure an interval of about 0.8 seconds which is sufficient to permit the distant end of any trunk, which has just been released, to restore to normal. When the timing interval is complete the GUARD TIME relays release, the direction of selections is reversed, and selections are resumed.

Direction of Selection Control

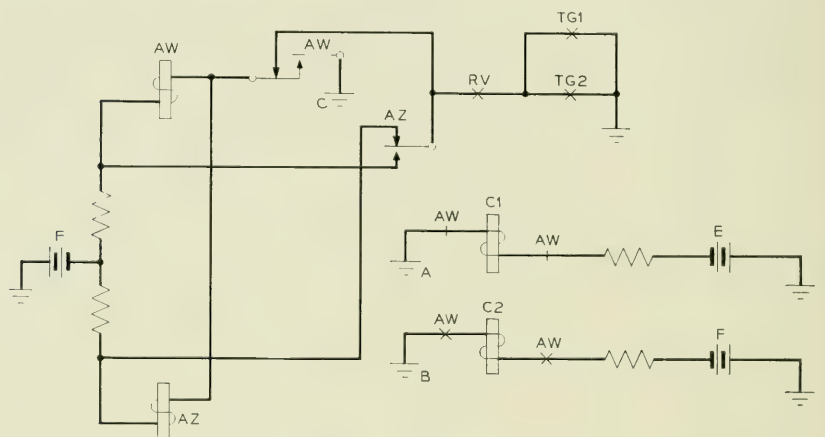
Two directions of selection are provided for both incoming and outgoing trunks to (1) by-pass incoming and out-going trunks which are in trouble, (2) to guard against blocking due to a single trouble in the controller itself. (See Fig. 12.) The directions of selection are controlled by two combinations of relays, plus three other relays controlled by these combinations. The ΔOW and ΔOZ relays control the outgoing directions of selection. With these relays in the unoperated condition the outgoing trunks will be selected in the low to high order and with them operated selections will be made in the high to low order. The outgoing trunks, in the normal course of events, are served starting with lowest or highest numbered trunks and proceeding to the highest or lowest numbered trunk depending upon the direction of selection at the time. When the last trunk has been selected the direction is reversed by either operating or releasing the ΔOW and ΔOZ relays depending upon the status quo ante. In case a short time-out is due to failure to close the crosspoint the direction of selection of outgoing trunks is reversed immediately. It is also reversed whenever a long time-out occurs, whatever the cause.

The ΔW and ΔZ relays form the combination which controls the incoming direction of selection. Controlled by them are the CONTROL relays c_1 and c_2 associated with the low to high and high to low directions respectively. With the ΔW relay unoperated the c_1 relay is operated and the direction of selection is from low to high. With the ΔW relay operated the c_1 relay is released, the c_2 relay is operated and the direction of selection is from high to low. The incoming direction of selection is reversed under the following conditions:

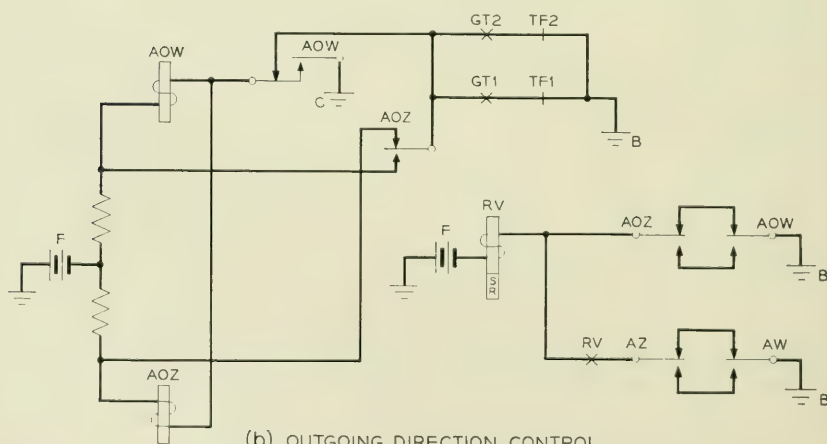
1. When the outgoing direction is reversed upon the last outgoing trunk being selected the incoming direction is also reversed if and when there are no more incoming trunks within the gate waiting to be served.
2. When a short time-out occurs due to the failure of the END relay to operate or failure of continuity check.
3. When a second failure to close crosspoint occurs.
4. When a long time-out occurs.

USE OF IDLE INDICATING LAMPS AT SWITCHBOARDS

A system of indicating an idle circuit in a trunk group is used at toll switchboards to relieve the operator of the necessity of making busy tests on the sleeve of the trunk. Associated with each trunk jack in the face of the switchboard is a lamp. The lamp associated with the first idle trunk in the group is normally lighted, while those lamps associated with the other trunks in the group remain dark even though those trunks are idle. When the trunk whose idle indicating lamp is lighted is seized, that idle indicating lamp is extinguished and the lamp associated with



(a) INCOMING DIRECTION CONTROL



(b) OUTGOING DIRECTION CONTROL

Fig. 12 — Direction of selection control.

the next idle trunk lights. The lamp does not necessarily move progressively thru the group but is always lighted on the lowest numbered idle trunk in the group.

The direct intertoll trunks and the trunks to a concentrating equip-

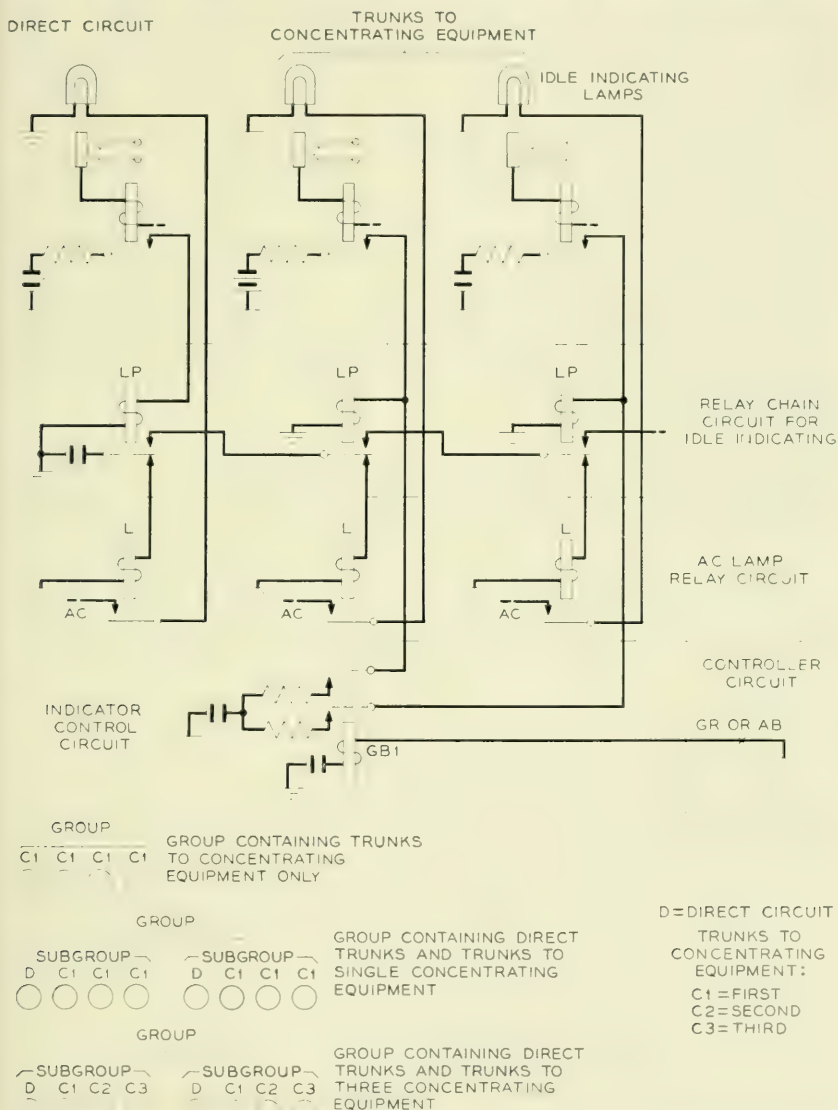


Fig. 13 — Idle trunk indicating control and some examples of groups containing direct trunks and trunks to concentrating equipment.

ment to the same destination form a single group of trunks in the switchboard multiple.

The use of idle trunk indicating lamps is required with such a group for the following reasons:

1. To force the traffic onto the direct trunks, if these are provided, when one or more of these trunks is idle. The direct circuits are first choice and calls should be routed to them when available to prevent overloading the outgoing trunks from the concentrator which might block traffic from other switchboards.

2. To make all trunks to the concentrator busy to operators when all intertoll trunks outgoing from the concentrator are in use. Fig. 13 is a simplified schematic of the idle indicating chain used in an outward switchboard for a group of trunks containing both direct and concentrator trunks. The arrangement shown is for a group consisting of direct trunks and trunks to a single concentrator. The idle indicating facilities are flexible and any desired grouping arrangement is attainable. A few of these arrangements are indicated by showing the relations of the jacks in the switchboard multiple.

USE OF THE INTERTOLL TRUNK CONCENTRATING EQUIPMENT IN AN OPERATOR ALTERNATE ROUTING ARRANGEMENT

The intertoll trunk concentrating equipment as previously stated is an independent system and its use and location are therefore flexible.

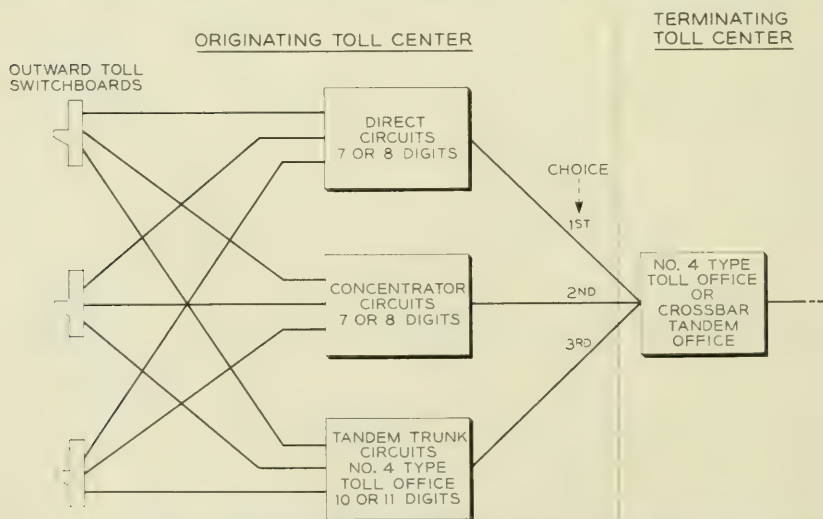


Fig. 14 — Use of concentrating equipment in an operator alternate routing arrangement.

However, it should be remembered that it contains a single controller which could be disabled by compound trouble, a condition which could be serious if the concentrating equipment were the sole means of handling traffic from any switchboard. It is intended that the concentrating equipment be used in conjunction with direct intertoll circuits and/or tandem trunks to a No. 4 type toll crossbar switching system. When used with direct circuits an operator alternate routing system ensues. Referring back to the paragraph which dealt with idle trunk indicating, it is noted that the direct trunks and the concentrator trunks form a single group or subgroup of trunks to a common destination. The direct trunks appear in the multiple at the head end of the group and are followed by concentrator trunks. The idle indicating lamps direct the operator to a direct trunk, when available, as first choice, and automatically direct the operator to a concentrator trunk when the direct trunks are in use. If the operator also has access to trunks to the toll crossbar office these trunks become third choice for use when the direct and concentrator trunks are busy. Fig. 14 illustrates this situation. The concentrating equipment may be located apart from the central toll building without losing the advantages of the alternate routing discussed above. Thus it is available for dispersing the toll plant to minimize the effect of disaster.

ACKNOWLEDGMENT

The development of the Intertoll Trunk Concentrating Equipment was the combined effort of many people. Important contributions were made by M. Posin and W. L. Shafer, Jr.

The Transistor as a Network Element*

By J. T. BANGERT

(Manuscript received October 7, 1953)

The development of the transistor has provided an active element having important advantages in space and power. As a result, the question arises whether strategic insertion of such active elements in passive networks might lead to interesting results. This paper gives a theoretical analysis confirmed by experiment, of certain possible network applications of transistors. Four general areas are considered in which transistors are used as follows: to reduce the detrimental effects of dissipative reactive elements, to eliminate the necessity for inductors in frequency selective circuits, to produce two terminal envelope delay structures having zero loss, and to invert the impedance of reactive structures. The conclusion is drawn that judicious interspersing of transistors in a transmission network enables performance to be achieved which would otherwise be unobtainable or uneconomical.

INTRODUCTION

It has become customary through the years to classify linear circuits as either active or passive. This convenient, but arbitrary, division has encouraged a philosophy that regards each as a separate and distinct domain. The recent spectacular advances in active devices suggest that in some cases the traditional boundaries should be erased and that a unified approach should be made.

In particular the development of the transistor offers the possibility of interspersing small active elements throughout a passive network to achieve certain desirable effects. This paper intends to survey a few of the ways in which a transistor can be used to advantage in transmission networks. The discussion is divided into four parts as follows:

1. Reduction of dissipation.
2. Elimination of inductance.
3. Production of delay.
4. Inversion of impedance.

* Presented in part at the Radio Fall Meeting, Toronto, Ontario, Oct. 28, 1953.

The first portion offers a new approach to the everpresent problem of imperfect reactive elements. The second portion discusses a method of combining resistance and capacitance with transistors to produce characteristics conventionally realized by inductance and capacitance. The third portion proposes a technique for obtaining any specified delay characteristic with a two terminal active structure. The fourth portion considers means of using a transistor to transform passive elements of ordinary size into passive elements of greatly reduced size.

I. REDUCTION OF DISSIPATION

For simplicity in the treatment of network problems it is frequently assumed that purely reactive elements will be used. In many cases this approximation is satisfactory; other times it is worthless, and a more realistic analysis must be made. In this latter case one possibility is to nullify the unwanted dissipation by means of a bridge balance.¹ This will entail the acceptance of some flat loss. Another possibility is to insert active elements within the network in order to supply just enough energy to offset the inherent dissipation of the elements. This second approach, until now relatively unexplored, is being tried with promising results. To avoid introducing new terminology the discussion will employ the concept of negative resistance which has been studied with interest by many investigators.²⁻¹⁸

Negative Resistance

Negative resistance is a misleadingly simple name applied to a complex phenomenon. The term implies behavior in some opposite sense to that of an ordinary positive resistance. This is true only for a limited range of frequencies and signal levels. As generally used negative resistance refers to a two terminal active network or electronic device in which the voltage-current ratio has a negative real part and negligible imaginary part.

TABLE I — NEGATIVE RESISTANCE

Parameter	Shunt Type	Series Type
Independent variable	Voltage controlled	Current controlled
Required external impedance	Short circuit stable	Open circuit stable
Effect of internal gain reduction	Increased magnitude of R_N	Decreased magnitude of R_N
Associated reactance	Parallel capacitance	Series inductance

For convenience the simple forms of negative resistance may be divided into two general classes which are duals in a network sense. Since these classes have been identified in the literature in several different ways, it seems desirable to summarize the major characteristics in the form given in Table I.

Therefore a shunt negative resistance is one whose magnitude is controlled mainly by the voltage across its terminals. It is short circuit stable which means it must operate into a low impedance. When the internal gain used to produce the effect is reduced, the magnitude of a shunt negative resistance increases. In addition it should be associated with a parallel capacitance to predict its behavior outside the working band of frequencies.

One method of producing a two terminal shunt negative resistance is to arrange a transistor as shown in Fig. 1(a). To facilitate prediction of the behavior of this combination it is desirable to derive an equivalent circuit.

Equivalent Circuit of a Transistor Shunt Negative Resistance

An equivalent circuit of the transistor and its associated network is shown in Fig. 1(b). By denoting each condenser reactance as jX , the circuit determinant, Δ , can be written as follows.

$$\begin{vmatrix} j2X & -jX & -jX & 0 \\ -jX & r_b + r_e + R_f + jX & -R_f & -r_e \\ -jX & -R_f & R_f + R_a + jX & -R_a \\ 0 & r_m - r_e & -R_a & r_e + r_c - r_m + R_a \end{vmatrix}$$

Next the input impedance is determined as Δ/Δ_{11} .

From this formula the exact general expression for the input impedance is found to be very cumbersome and will not be given. A useful approximation can be found by making some simplifying assumptions

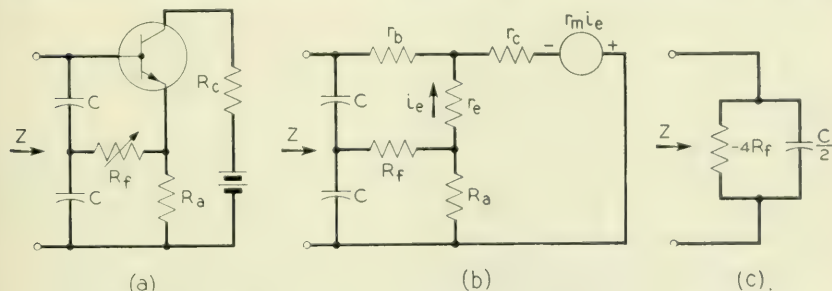


Fig. 1 — Equivalent circuit of transistor shunt negative resistance.

as follows:

$$\begin{array}{ll} \text{Let } r_c > > r_b & r_m > > r_b \\ r_c > > r_e + R_a & r_m > > r_e + R_a \end{array}$$

Under these conditions the network can be represented by the equivalent circuit shown in Fig. 1(c). This circuit consists of a parallel combination of resistance and capacitance in which the capacitance is that of the original two capacitances in series and the resistance is negative and equal to four times the feedback resistor, R_f . Hence the magnitude of the generated shunt negative resistance can be controlled by adjustment of R_f . One measure of the accuracy of this approximation is how much the "constants" of the equivalent circuit change with frequency. Calculations of a typical case show that deviations in frequency of ± 5 per cent cause deviations in both capacitance and negative resistance of ± 0.05 per cent. Hence this approximation is very accurate for narrow band applications.

This circuit can now be used to advantage in a band filter.

Confluent Band Filter

A conventional confluent band filter is shown in Fig. 2(a). In this structure the presence of dissipation in the series branches impairs performance by introducing flat loss, whereas any dissipation in the shunt branch not only produces flat loss, but, worse still, causes rounding of the transmission characteristic at the edges of the band. For narrow filters having a small percentage band width, any appreciable dissipation in the shunt arm can degrade the transmission characteristic beyond a reasonable tolerance. One good answer to this problem is to use elements having an extremely low resistive component such as quartz crystals. However, quartz is expensive and has other limitations. Another solution is to build a negative resistance into the filter so as to reduce the inherent element dissipation to zero or at least to a tolerable value. In the present case a shunt negative resistance will be used to compensate the shunt branch. This is done by splitting the shunt capacitance of Fig. 2(a) and inserting the circuit of Fig. 1(a). This can be illustrated by an example. When the filter of Fig. 2(a) is designed to give a 5 per cent band at a midfrequency of 10 kc and impedance level of 600 ohms the shunt branch offers an undesirably low impedance to the compensating transistor circuit and in addition requires cumbersome element values. Both difficulties can be corrected by using capacitive impedance transformations on each side of the shunt branch, thereby

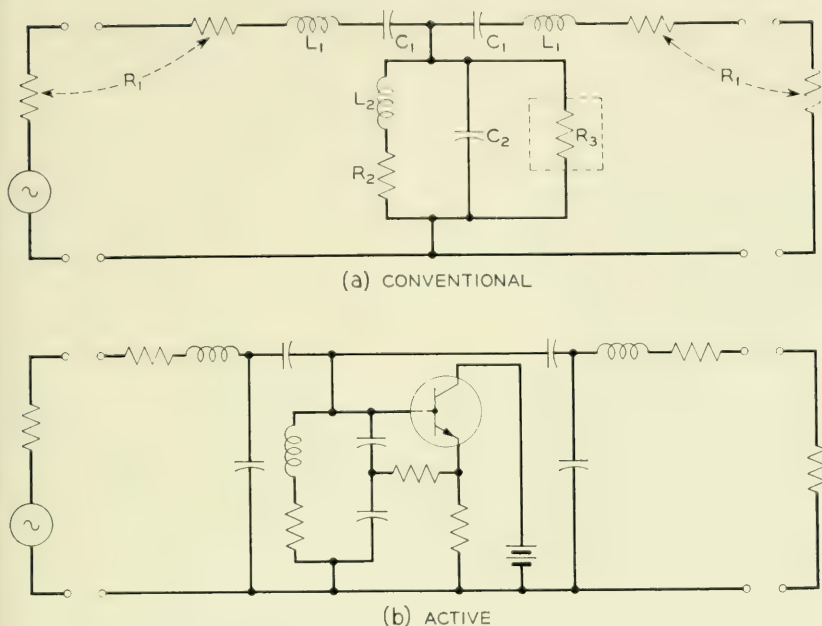


Fig. 2 — Confluent band filters. (a) Conventional. (b) Active.

raising the impedance of the shunt branch without changing the impedance level at the input and output terminals. At the same time the elements assume much more reasonable values. The modified configuration together with the active portion is shown in Fig. 2(b).

Using the filter described above, a series of transmission curves were calculated and are shown in Fig. 3. When ideal elements are assumed, the transmission is

$$e^{-\theta} = \frac{k^3 p^3}{(p^2 + kp + 1)[p^4 + kp^3 + (k^2 + 2)p^2 + kp + 1]}$$

$$\text{where } k = \frac{f_2 - f_1}{\sqrt{f_2 f_1}} \quad \text{and} \quad p = j\omega$$

Calculation of this expression results in the classical characteristic labeled "Ideal Passive". When, however, typical values of element resistance are introduced, the transmission is

$$e^{-\theta} = \frac{2R_1(\delta_1 p + \delta_2 p^2)(\beta_1 p + \beta_2 p^2 + \beta_3 p^3)}{(\alpha_0 + \alpha_1 p + \alpha_2 p^2 + \alpha_3 p^3 + \alpha_4 p^4)^2 - (\delta_1 p + \delta_2 p^2)^2}$$

where

$$\alpha_0 = R_1(1 + g)$$

$$\alpha_1 = R_1[T_1(1 + g) + T_2 + T_{23}] + R_2T_1$$

$$\alpha_2 = R_1[A_1(1 + g) + T_1(T_2 + T_{23}) + A_2] + R_2T_1T_{22}$$

$$\alpha_3 = R_1[A_1(T_2 + T_{23}) + A_2T_1]$$

$$\alpha_4 = R_1A_1A_2$$

$$\delta_1 = R_2T_1$$

$$\delta_2 = R_2T_1T_{22}$$

$$\beta_1 = T_1(1 + g)$$

$$\beta_2 = T_1(T_2 + T_{23})$$

$$\beta_3 = T_1A_2$$

$$g = \frac{R_2}{R_3}$$

$$T_1 = R_1C_1$$

$$T_2 = R_2C_2$$

$$T_{22} = \frac{L_2}{R_2}$$

$$T_{23} = \frac{L_2}{R_3}$$

$$A_1 = L_1C_1$$

$$A_2 = L_2C_2$$

This expression with the compensating negative resistance $R_3 = \infty$ produces the characteristic labeled "Practical Passive".

The dotted curves show the effect of adding various amounts of negative resistance to the shunt branch by letting R_2 assume negative values. The number on each dotted curve is the ratio of the resistive component of the shunt arm at anti-resonance to the magnitude of the compensating negative resistance. For example, for the curve labeled $\rho = 1$, the resistance in the shunt arm is entirely compensated so that the loss is only that due to the resistance in the series arms. By increasing the amount of compensation in the shunt arm so that $\rho > 1$, called overcompensation it is possible effectively to nullify the losses in the series arm as well. Comparison of the active curve labeled $\rho = 1.21$ with the "ideal passive" curve shows that this technique of resistance compensation can produce a practical filter having a characteristic equal to that of a filter having ideal elements. Filters of this type have already been successfully used in field test equipment.

When the degree of compensation is increased still further, the filter begins to provide gain in the band as shown by the $\rho = 1.36$ curve. It is clear that continued increases in the compensation will eventually absorb the terminations causing the structure to become unstable.

Although the curves given in Fig. 3 are all calculated, tests on experimental models show excellent agreement. To a reader having long experience with passive filters the development of negative insertion loss may seem a little surprising. In order to lend an air of authenticity to this midband gain it is instructive to consider the behavior of a resistive tee section having a negative element. This is a reasonable analogue,

because at midfrequency the reactive component becomes zero in each branch of the filter.

Transmission of Symmetrical Tee

The insertion loss of a symmetrical tee section with positive resistance is a well known concept. It is doubtful, however, if the behavior of a tee with a negative element is equally well known. Consider the section shown in Fig. 4 operating between terminations R and having series arms, R_A and a shunt arm, R_B . Normalize by letting $a = R_A/R$ and $b = R_B/R$. Insertion loss is plotted vs. b with a as the third parameter. For b positive the usual loss pattern results; for b negative, a more complex situation develops. When b is very large and negative, the section is still producing a small loss, but as b becomes smaller in magnitude the loss drops to zero and finally becomes a gain. There is a lower limit on the magnitude of b beyond which oscillations will occur. This limit is reached when $2b = -(a + 1)$.

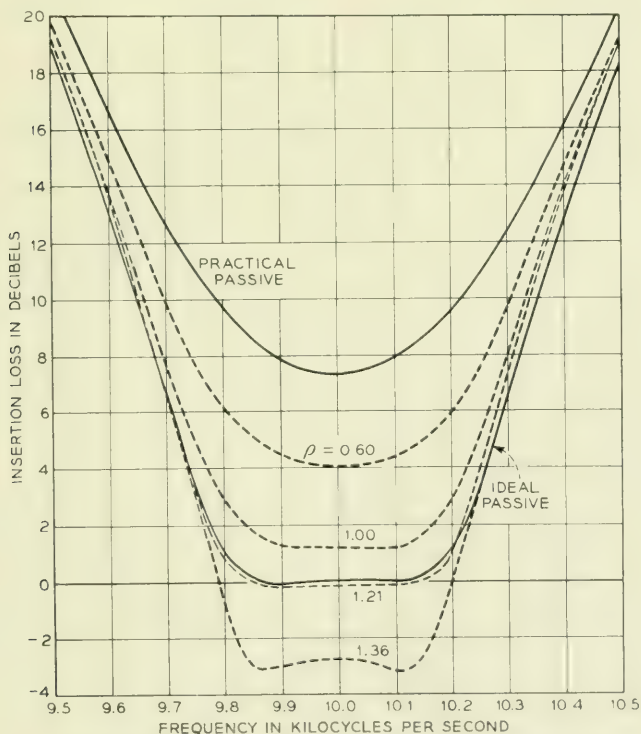


Fig. 3 — Transmission of confluent band filters.

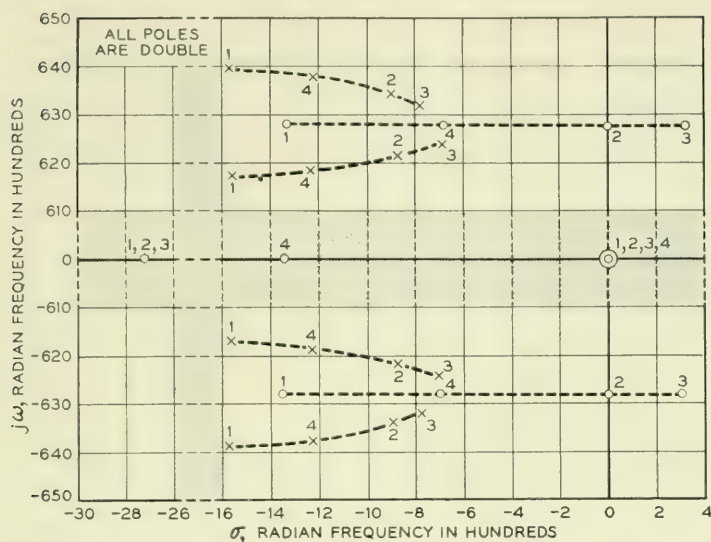


Fig. 5 — Effect of reducing dissipation on singularities of confluent band filter.

the real frequency axis is a function of the amount of dissipation in the elements. When a value of negative resistance corresponding to the $\rho = 1$ curve in Fig. 3 is added to the passive filter, the singularities move from positions marked 1 to those marked 2 in Fig. 5. Adding a larger amount of negative resistance corresponding to the $\rho = 1.21$ curve in Fig. 3 produces the singularities marked 3. It should be noted that the infinite loss point on the negative sigma axis as well as the two at the origin have not moved. If the dissipation in the shunt branch is reduced by removing the coil and replacing by one having half as much resistance, the singularities change from position one to position four. In this case the infinite loss point on the sigma axis does move. This illustrates that the change in pattern of singularities resulting from use of negative resistance is similar to, but not the same as, that resulting from use of passive inductors having higher values of Q .

M-Derived Band Pass

In order to provide a sharp cut-off in a filter use is often made of m -derived peak sections. In the configuration shown in Fig. 6 loss peaks will occur at selected frequencies above and below the pass band provided the elements are nearly free of dissipation. The closer the attenuation peaks are to the pass band the more nearly free from dissipation the

elements must be for good performance. As in the previous case, a transistor negative resistance is used to compensate the anti-resonant portion of the shunt arm. The magnitude of this resistance can also be adjusted to serve the additional purpose of compensating for resistance in the series resonant circuit in the shunt arm, as well as the series resonant circuits in the series arms.

The transmission of a non-dissipative m -derived band filter between unit resistive terminations is

$$e^{-\theta} = \frac{kp[(1 - m^2)p^4 + (2 - 2m^2 + k^2)p^2 + (1 - m^2)]}{(mp^2 + kp + m)[p_4 + kmp^3 + (k^2 + 2)p^2 + kmp + 1]}$$

$$\text{where } m = \sqrt{1 - \left(\frac{f_c}{f_\infty}\right)^2}$$

Assuming no dissipation a peak section with an m of 0.86 will give the characteristic shown in Fig. 7 labeled "Ideal Passive". However, when this filter is constructed with typical elements the curve labeled "Practical Passive" results. By introducing a suitable amount of negative resistance the transmission of the practical filter can be made comparable to that of the ideal filter, as illustrated by the curve labeled "Practical Active".

For maximum utility active filter sections must be capable of being connected in tandem to form composite filters without instability, reflections, or interactions. Fig. 8 shows that these filters meet this requirement by giving the measured transmission of a band filter composed of two dissimilar peak sections. On the basis of attainable electrical charac-

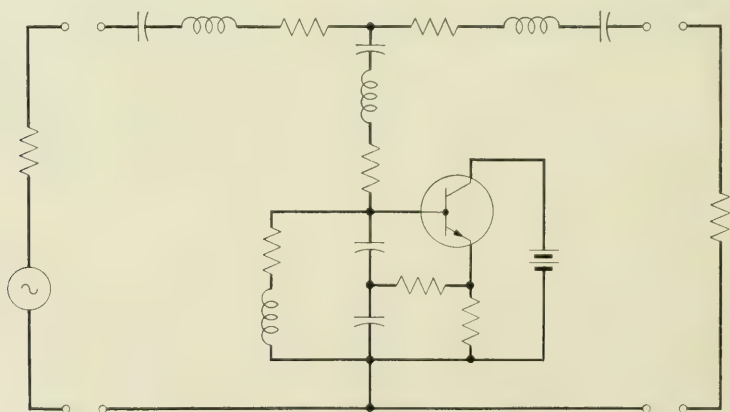


Fig. 6 — Active M-derived band filter.

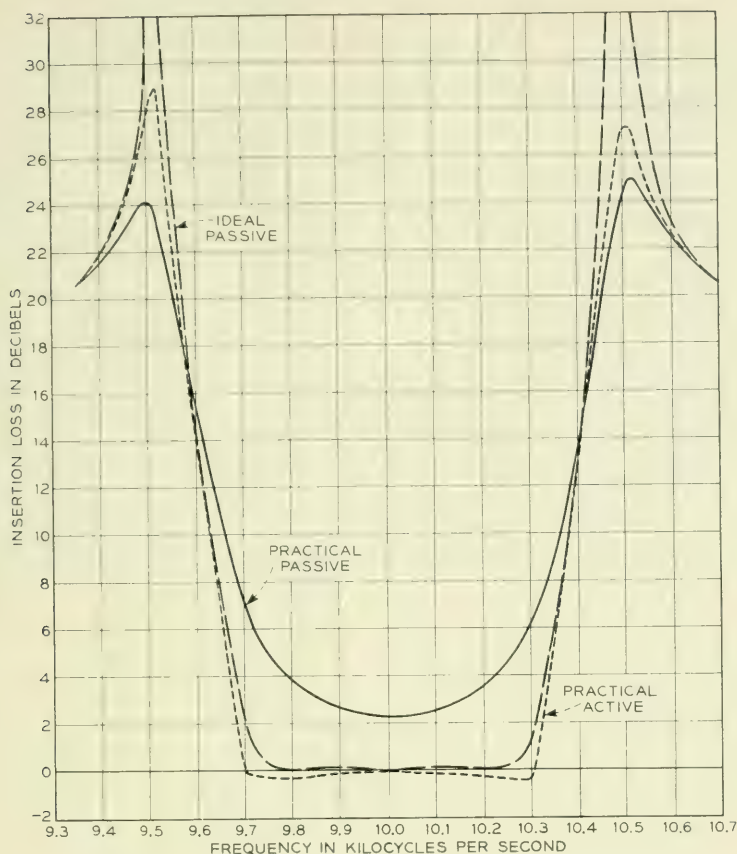


Fig. 7 — Transmission of M-derived band filters.

teristics, active filters of this kind appear to offer potential competition to the crystal channel filters used in broad band carrier systems.

Working Model

To further emphasize this fact the photograph of Fig. 9 shows a model of a composite band filter designed to transmit a 4-ke band at a midfrequency of 98 kc. This model contains seven miniature, adjustable, ferrite inductors, miniature capacitors, and two *n-p-n* junction transistors. The transmission characteristic is shown in Fig. 10. Hence in some cases by employing active circuitry it is possible to use miniature components thereby gaining at least an order of magnitude in the size and weight of structure.

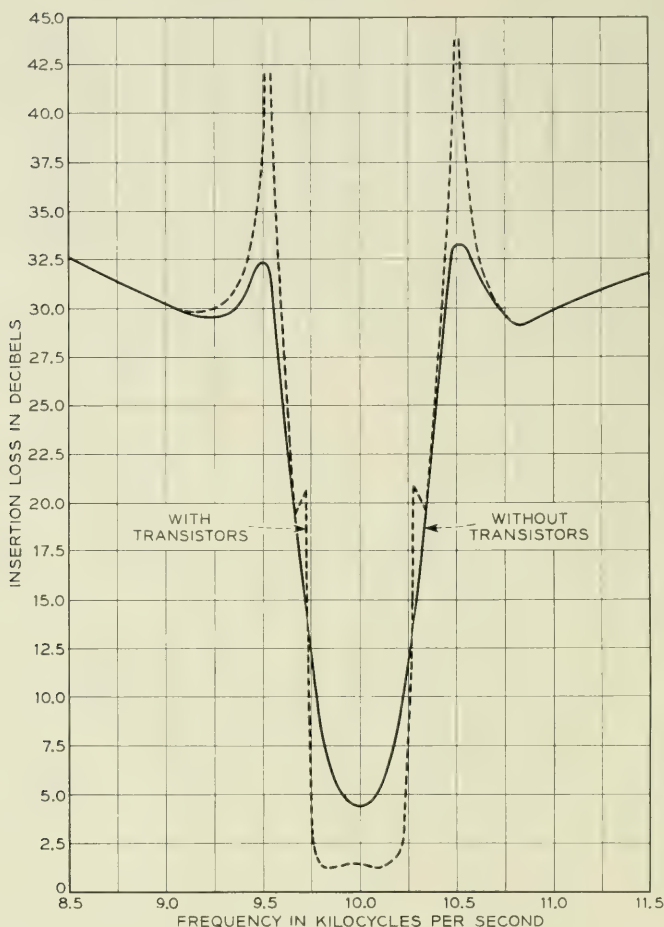


Fig. 8 — Resistance compensation of multi-section filter.

Series Negative Resistance

For satisfactory performance in many applications a series resonant circuit should approach zero impedance at the resonant frequency. To reduce the residual dissipation in an ordinary tuned circuit a series negative resistance, consisting of two transistors, can be used. This technique is illustrated in Fig. 11 which shows how a purely reactive shunt branch can be achieved in either an m -derived low pass filter or a confluent band elimination filter.

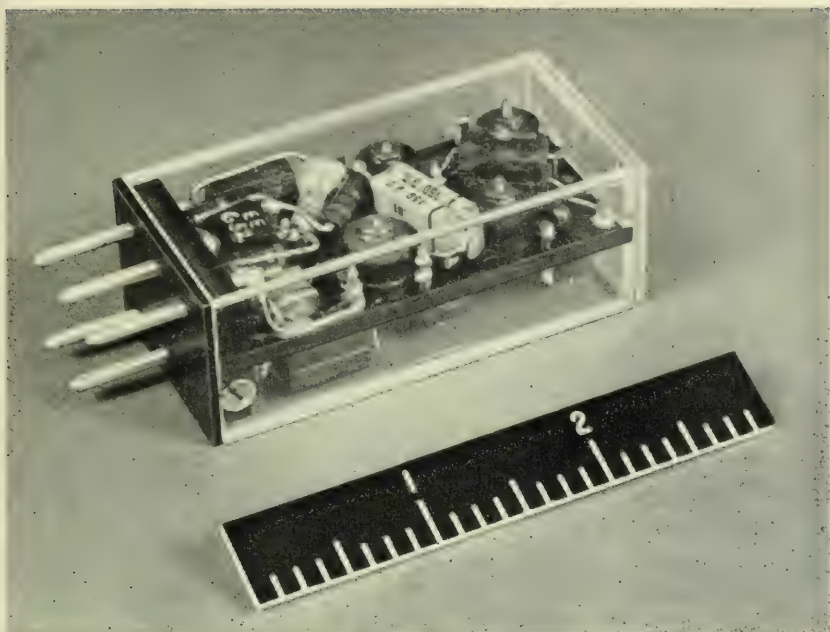


Fig. 9 — 98-kc active channel filter.

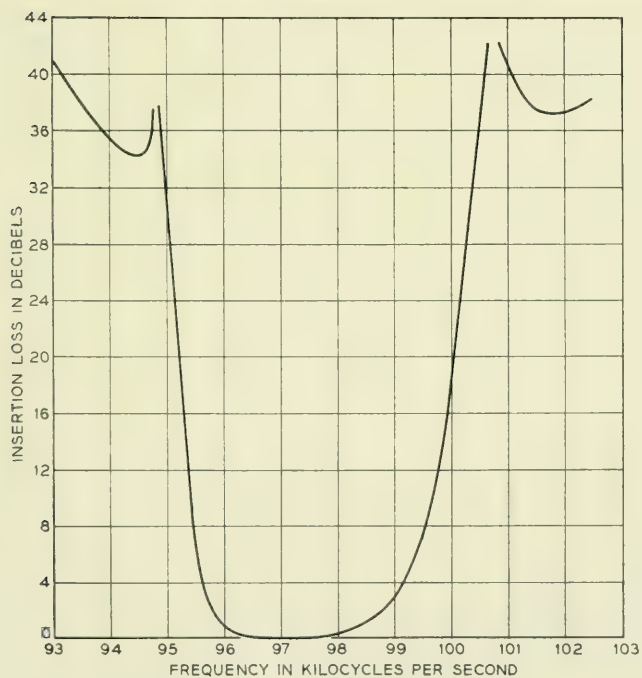


Fig. 10 — Active channel filter.

II. ELIMINATION OF INDUCTANCE

In the practical realization of frequency selective networks it is sometimes awkward, difficult, or even impossible to make effective use of coils as inductive elements. This is true, because of severe limitations on space, exacting tolerances on undesired modulation, or necessity for operation at extremely low frequencies.

It has been well known for some time that inductive elements can be eliminated without restricting the repertoire of the network designer provided he is willing to purchase this freedom by introducing active elements to supply gain.^{19, 20}

It can be easily shown that the transmission through a high gain feedback amplifier is proportional to the product of the short circuit transfer admittance of the input network and the short circuit transfer impedance of the feedback network:

$$e^{-\theta} = Y_i Z_f$$

In addition it is also known from energy relations that passive networks containing only one kind of reactance cannot produce complex poles in the short circuit transfer admittance. It is instructive to consider the application of these principles to some familiar kinds of transmission networks. These networks can be logically divided into two classes: those which are primarily concerned with amplitude such as filters, and those mainly concerned with phase such as delay equalizers.

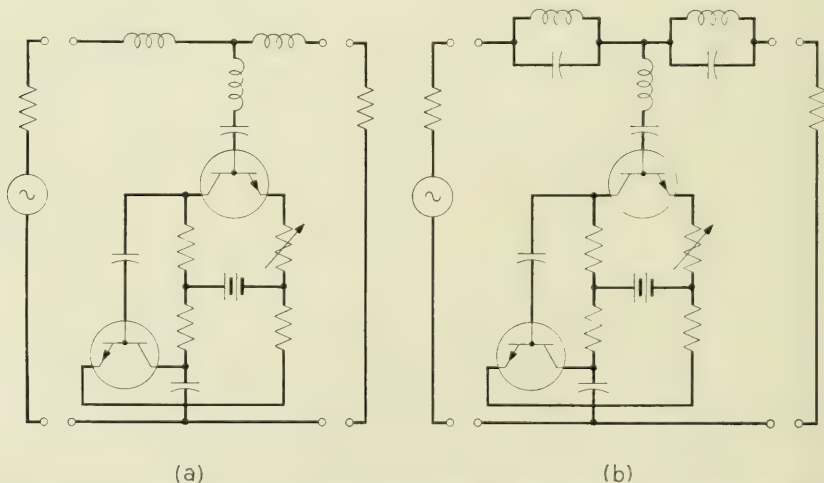


Fig. 11.— Use of series negative resistance. (a) M-derived low pass filter. (b) Band elimination filter.

*Non-inductive Filters**Low Pass*

Consider first an image parameter, constant- k , low pass filter which is usually built as a ladder-type structure of series inductance and shunt capacitance. A full section contains three reactive arms and produces an asymptotic loss that increases 18 db per octave.

The transmission is given by the following expression

$$e^{-\theta} = \frac{1}{(1 + \omega_0^{-1}p + \omega_0^{-2}p^2)(1 + \omega_0^{-1}p)}$$

where ω_0 = cut-off frequency in radians per second.

The function has three poles, one real and two complex conjugate. The question now arises how this function can be divided between the input and feedback networks so as to be physically realizable. Since we know that a passive R - C structure cannot have complex poles in the short circuit transfer admittance, there is no choice but to use the impedance function in the feedback circuit for this purpose. It is now found that any R - C structure which will provide the complex poles insists on providing a real zero for good measure. This unwelcome zero can be nullified by supplying its counterpart as a pole in the admittance function. The transmission is now rewritten, as follows:

$$e^{-\theta} = \left[\frac{1}{(1 + \omega_0^{-1}p)(1 + ap)} \right] \left[\frac{1 + ap}{1 + \omega_0^{-1}p + \omega_0^{-2}p^2} \right]$$

and the singularities are shown in Fig. 12(a). Since the original transmission function also requires a real pole, the admittance function must now supply two real poles. A simple ladder structure having three series resistances and two shunt capacitances meets this requirement. The complex poles cannot be supplied by a ladder structure, but require some sort of bridge such as shown in Fig. 12(a).

At low frequencies the transmission through the filter depends on the ratio of the total series input resistance to the total resistance in the bridge arm of the feedback network. Therefore any amount of flat loss or a moderate flat gain through the filter can be obtained simply by adjusting the ratio of impedance levels of the input and feedback networks.

Simulation of functions by this technique does not provide a unique solution since there is considerable freedom in choice of configuration and location of the cancelling pole and zero.

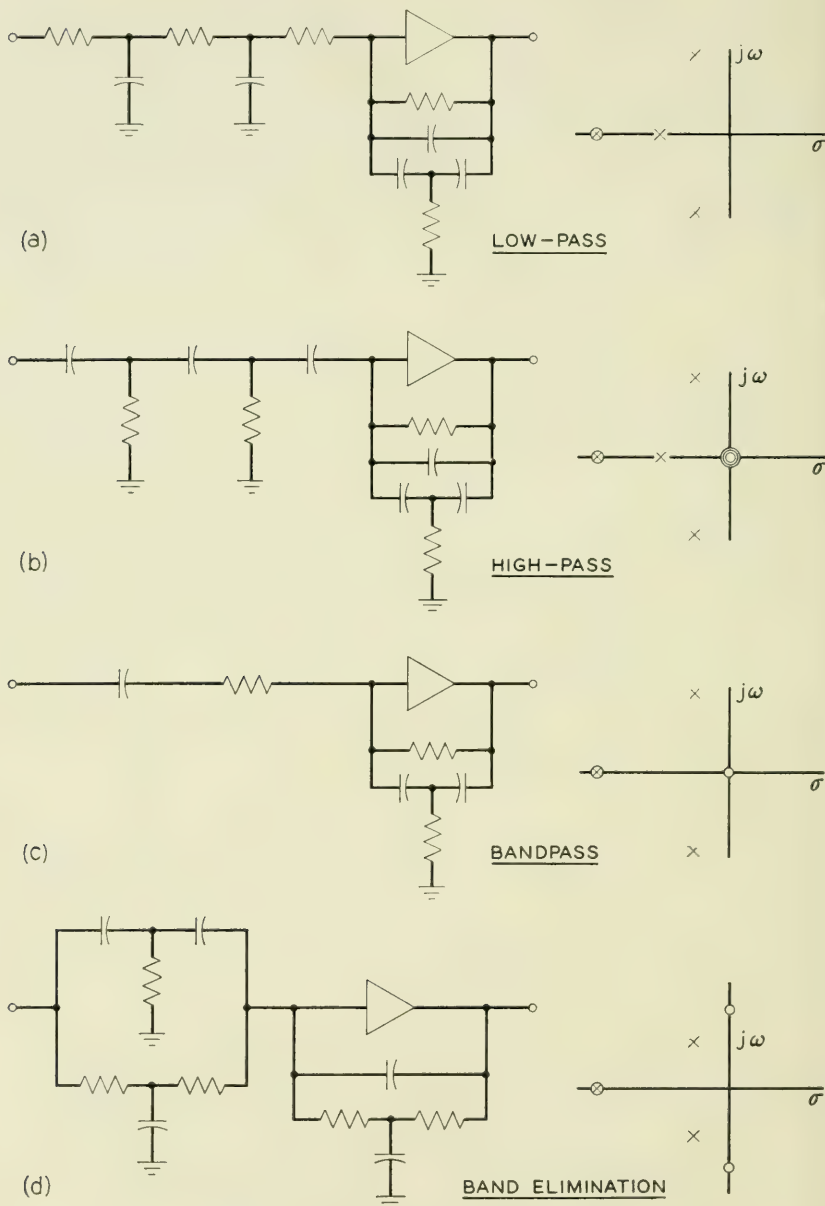


Fig. 12 — Non-inductive active filters.

High Pass

Consider next a high pass filter with cut-off at ω_0 . The transmission is

$$e^{-\theta} = \frac{\omega_0^{-3} p^3}{(1 + \omega_0^{-1} p + \omega_0^{-2} p^2)(1 + \omega_0^{-1} p)} \\ = \left[\frac{\omega_0^{-3} p^3}{(1 + \omega_0^{-1} p)(1 + ap)} \right] \left[\frac{1 + ap}{1 + \omega_0^{-1} p + \omega_0^{-2} p^2} \right]$$

The complex plane plot in Fig. 12(b) shows exactly the same pattern of singularities as the low pass case with the addition of three zeros at the origin. To realize this function the feedback network remains unchanged, whereas the input network becomes a ladder in which the positions of the resistances and capacitances are interchanged.

Band Pass

A series resonant branch inserted in series between resistive terminations is a simple form of band pass filter having the following transmission:

$$e^{-\theta} = \frac{\omega_m^{-1} Q^{-1} p}{1 + \omega_m^{-1} Q^{-1} p + \omega_m^{-2} p^2} = \left[\frac{\omega_m^{-1} Q^{-1} p}{1 + ap} \right] \left[\frac{1 + ap}{1 + \omega_m^{-1} Q^{-1} p + \omega_m^{-2} p^2} \right]$$

where ω_m is the radian frequency of the peak and Q is a measure of the sharpness of the peak.

The singularities shown in Fig. 12(c) consist of a zero at the origin and two complex conjugate poles. Once again the complex poles are obtained by a bridge circuit in the feedback path. The usual penalty is incurred by the appearance of a real zero which must be cancelled by a real pole. Therefore the admittance function must supply a zero at the origin and one real pole. This is done by a series combination of resistance and capacitance in the input circuit.

Band Elimination

A parallel resonant branch inserted in series between resistive terminations is a simple form of band elimination filter having the following transmission:

$$e^{-\theta} = \frac{1 + \omega_m^{-2} p^2}{1 + \omega_m^{-1} Q p + \omega_m^{-2} p^2} = \left[\frac{1}{1 + ap} + \frac{\omega_m^{-2} p^2}{1 + ap} \right] \left[\frac{1 + ap}{1 + \omega_m^{-1} Q p + \omega_m^{-2} p^2} \right]$$

The singularities consist of two conjugate zeros on the real frequency axis and two complex conjugate poles. A bridge circuit in the feedback

path supplies the complex conjugate poles and a parasitic real zero, while a parallel tee in the input path provides the conjugate zeros and a real pole.

It has been found that the experimental performance of the various non-inductive filters described can be predicted with precision from the theory.

Non-inductive Phase Sections

Non-minimum phase networks are used extensively to provide a specified variation in phase with frequency without introducing any change in attenuation. Such all-pass networks consisting only of reactive elements are usually designed as lattices or bridged tee sections. It is theoretically possible and practically desirable to represent any complex all-pass structure by tandem arrangements of two basic all-pass sections called first degree and second degree. The first degree structure pro-

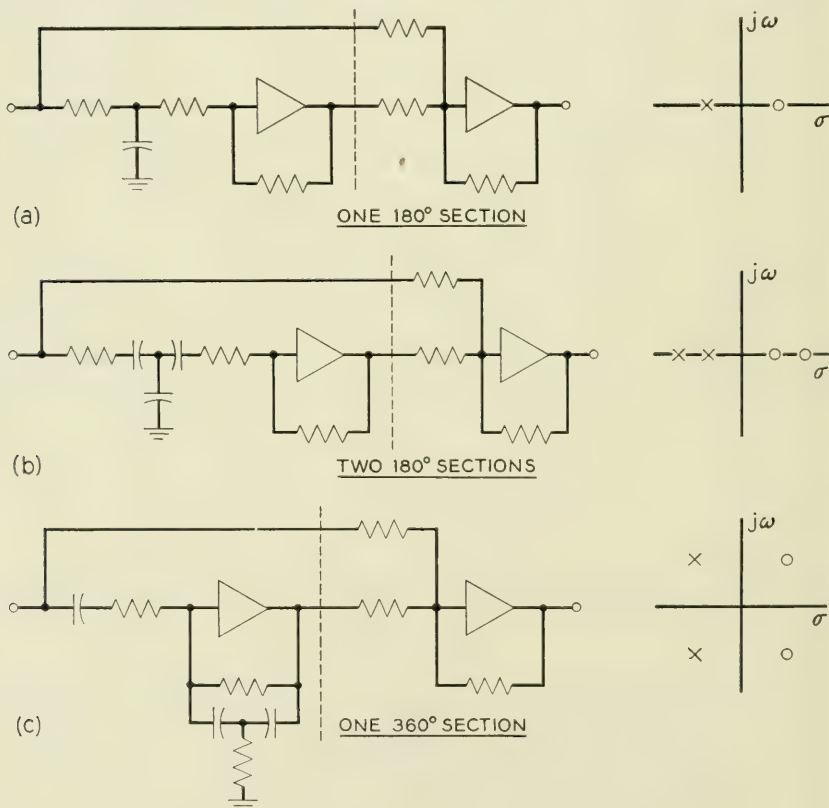


Fig. 13 — Non-inductive active phase sections.

vides a total change in phase of 180° and is characterized by a single pole-zero pair symmetrically located on the σ axis as shown in Fig. 13(a). Only one parameter, the distance from the origin can be chosen. The second degree structure provides a maximum phase shift of 360° and is characterized by two conjugate poles in the left half plane and two symmetrically located zeros in the right half plane shown in Fig. 13(c). The two parameters which can be selected are the rectangular coordinates of one singularity.

It has been suggested that these functions can be realized without benefit of inductance. Here again numerous arrangements are possible, but only a few examples will be given. The basic operation is to perform one division on the original transmission function resulting in a quotient of unity and a fractional remainder of opposite sign. The fractional remainder is then synthesized by a RC network in conjunction with an amplifier.

Single 180° Section

The transmission of a single 180° section is

$$e^{-\theta} = \frac{1 - \omega_0^{-1}p}{1 + \omega_0^{-1}p} = \frac{2\omega_0}{\omega_0 + p} - 1$$

In this case the fractional remainder consists of only one real pole which is realized by the R - C structure shown in Fig. 13(a).

Two 180° Sections

The overall transmission of two 180° sections in tandem is the product of each transmission

$$e^{-\theta} = -\left[\frac{1 - \omega_1^{-1}p}{1 + \omega_1^{-1}p}\right]\left[\frac{1 - \omega_2^{-1}p}{1 + \omega_2^{-1}p}\right] = \frac{2(\omega_1^{-1} + \omega_2^{-1})p}{(1 + \omega_1^{-1}p)(1 + \omega_2^{-1}p)} - 1$$

In this case the fractional remainder consists of one real zero and two real poles which are realized by the R - C structure shown in Fig. 13(b).

Single 360° Section

By far the most common phase corrector is the 360° section whose transmission is

$$\begin{aligned} e^{-\theta} &= -\left[\frac{1 - Q^{-1}\omega_m^{-1}p + \omega_m^{-2}p^2}{1 + Q^{-1}\omega_m^{-1}p + \omega_m^{-2}p^2}\right] \\ &= 2\left[\frac{Q^{-1}\omega_m^{-1}p}{1 + ap}\right]\left[1 + Q^{-1}\omega_m^{-1}p + \omega_m^{-2}p^2\right] - 1 \end{aligned}$$

In this case the fractional remainder consists of a zero at the origin and two conjugate complex poles which are realized by the R - C structure shown in Fig. 13(c).

III. PRODUCTION OF DELAY

It has also been proposed that a two terminal active delay equalizer can be constructed with the help of a negative resistance. As shown in Fig. 14 a two terminal network Z is connected between a resistive source and load, each of magnitude, one quarter R_0 . The network Z consists of a parallel combination of a reactive network jX and a negative resistance ($-R_0$). The transmission through Z is

$$e^{-\theta} = \frac{\frac{R_0}{2}}{\frac{R_0}{2} + Z} = \frac{\frac{R_0}{2}}{\frac{R_0}{2} + \frac{jR_0X}{R_0 - jX}} = \frac{R_0 - jX}{R_0 + jX}$$

This is the desired function, because the amplitude of the transmission

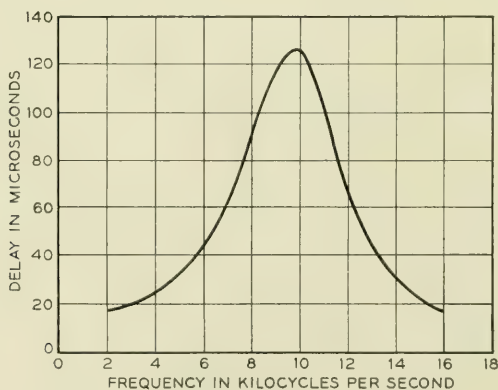
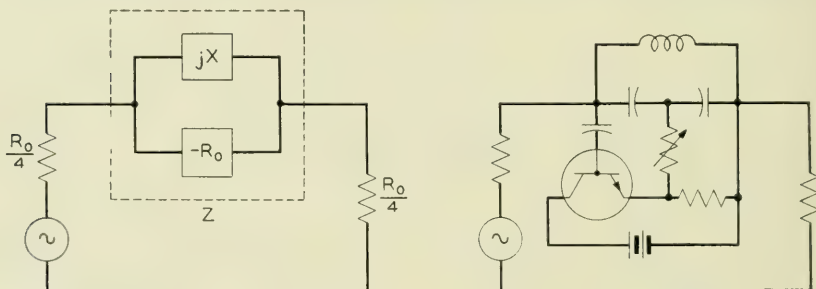


Fig. 14 — Active all-pass section.

is unity regardless of the size of X , and the phase and resultant delay are frequency dependent, because X is a function of frequency. It is theoretically possible to produce the most complicated delay equalizer characteristic by this method provided the negative resistance remains constant over the desired frequency band. As examples only a single 180° section and a single 360° section will be considered.

A 360° section results when the reactance is a single antiresonance given by

$$X = \frac{\omega L}{1 - \omega^2 LC}$$

the transmission is

$$e^{-\theta} = \frac{(p - k)^2 + \omega_m^2}{(p + k)^2 + \omega_m^2} = \frac{1 - Q^{-1}\omega_m^{-1}p + \omega_m^{-2}p^2}{1 + Q^{-1}\omega_m^{-1}p + \omega_m^{-2}p^2}$$

where

$$k = \frac{Q}{2} \omega_m \quad \text{and} \quad Q = (\omega_m R_0 C)^{-1}$$

In this case there are two degrees of freedom, namely, the width of the delay characteristic and the location of the peak frequency.

The circuit is shown on Fig. 14, where the transistor supplies the negative resistance, the magnitude of which is controlled by the adjustable resistance. A typical delay characteristic is also shown on Fig. 14.

A single 180° section can be obtained by simply omitting the coil in the above circuit. This is equivalent to letting $X = -(\omega C)^{-1}$ so that

$$e^{-\theta} = \frac{p - \omega_0}{p + \omega_0}$$

where $\omega_0 = (R_0 C)^{-1}$

IV. IMPEDANCE INVERSION

Two networks are said to be inverse if the product of their impedance functions is a constant. Given a network of passive elements, there are standard topological methods for finding its structural inverse if it exists. Another method is to use an active circuit in conjunction with the given impedance so that the combination offers an impedance inverse to that of the original impedance. This is a special case of modifying an impedance by feedback.²¹ By means of such methods passive circuit elements can be made to appear electrically much larger or much smaller

than they really are. For example, as shown in Fig. 15(a), by using an active element a parallel combination of convenient elements such as a 0.1 henry inductance and a 10,000 $\mu\mu\text{f}$ capacitance can be made to look like the series combination in Fig. 15(b) of difficult or sensitive elements like a 100 henry inductance and a 10 $\mu\mu\text{f}$ capacitance.

This transformation can be made with the transistor circuit of Fig. 15(c) which is also drawn as the equivalent circuit of Fig. 15(d).

In the circuit of Fig. 15(c) the resistor R_A is adjusted so that the

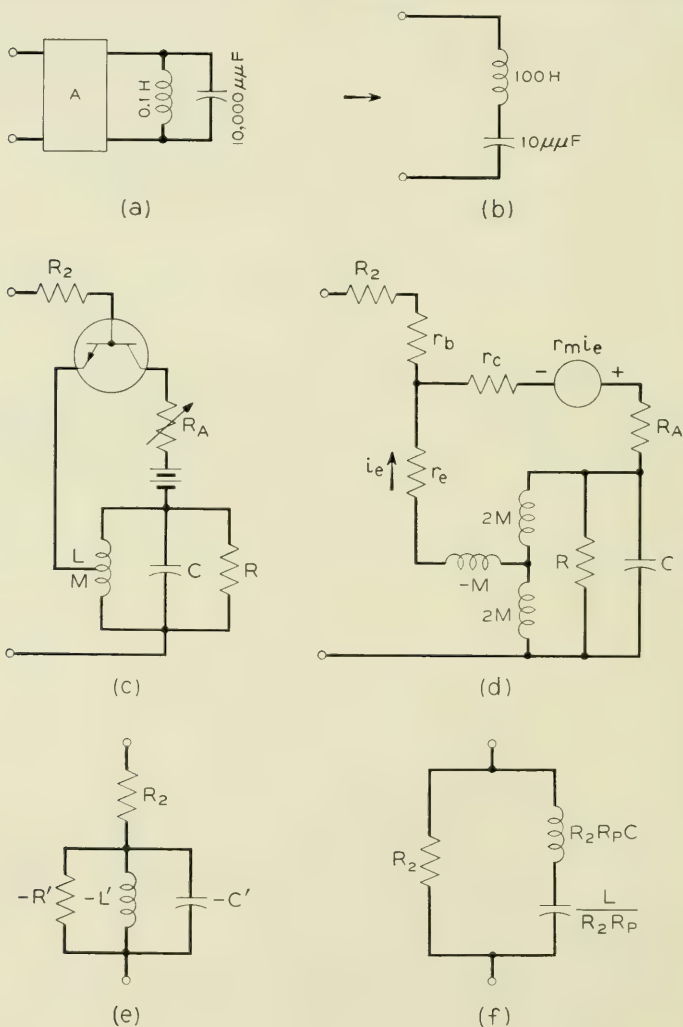


Fig. 15—Impedance inversion.

parameters of the transistor and the associated external circuit will satisfy the following equation

$$\frac{r_c}{2r_0} = \frac{2R_2}{R_p} + 1$$

To eliminate non-essentials it will be assumed that r_b and r_c are negligibly small, and $R_a \ll r_c$. Then by a straightforward, but lengthy, analysis the driving point impedance is found to be

$$Z = R_2 - \frac{pL'}{p^2L'C' + p\frac{L'}{R'} + 1}$$

$$\begin{aligned} \text{where } L' &= \lambda L & R_p &= \frac{4r_0R}{R + 4r_0} \\ R' &= \lambda R_p & \lambda &= \frac{r_c - 2r_0}{4r_0} \\ C' &= \frac{C}{\lambda} & r_0 &= r_c - r_m \end{aligned}$$

The circuit representing this impedance is shown in Fig. 15(e). Since negative elements are not convenient a final transformation is made to the circuit shown in Fig. 15(f).

CONCLUSION

The distinctive properties of the transistor suggest careful consideration of a philosophy which regards the transistor as a circuit element to be introduced at strategic points within a network. Initial work indicates that the judicious interspersion of transistors in a transmission network makes possible performance otherwise unobtainable or uneconomical. This paper has presented examples of how transistors may be used to reduce dissipation, to eliminate inductance, to produce delay, and to invert impedance. Undoubtedly this is only the beginning of exploration which should extend the horizons of network design.

ACKNOWLEDGEMENT

The author is indebted to many associates in Bell Telephone Laboratories, in particular to E. I. Green for basic philosophy and to W. R. Lundry for much helpful advice, and many suggestions, including the novel concepts of non-inductive phase sections and active delay equalizers.

BIBLIOGRAPHY

1. H. Bode, U. S. Patent 2,002,216, May 21, 1935.
2. A. Hull, Description of the Dynatron, *Proc. I.R.E.*, **6**, p. 5, Feb., 1918.
3. A. Bartlett, Boucherot's Constant Current Networks and their Relation to Electric Wave Filters, *J. I. E. E.*, **65**, p. 373, March, 1927.
4. H. Mouradian, Some Long Distance Transmission Problems, *Journal Franklin Inst.*, **207**, p. 165, Feb., 1929.
5. B. van der Pol, New Transformation in Alternating — Current Theory with Application to Theory of Audition, *Proc. I.R.E.*, **18**, p. 221, Feb., 1930.
6. L. Verman, Negative Circuit Constants, *Proc. I.R.E.*, **19**, p. 676, April, 1931.
7. G. Crisson, Negative Impedances and the Twin 21-Type Repeater, *B.S.T.J.*, **10**, p. 485, July, 1931.
8. F. Colebrook, Voltage Amplification with High Selectivity by Means of the Dynatron Circuit, *Wireless Eng.*, **10**, p. 69, Feb., 1933.
9. S. Cabot, Resistance Tuning, *Proc. I.R.E.*, **22**, p. 709, June, 1934.
10. E. Herold, Negative Resistance and Devices for Obtaining It, *Proc. I.R.E.*, **23**, p. 1201, Oct., 1935.
11. L. Curtis, Selectivity Control for Radio, U. S. Patent 2,033,330, March 10, 1936.
12. C. Brunetti, Clarification of Average Negative Resistance with Extension of its Use, *Proc. I.R.E.*, **25**, p. 1595, Dec., 1937.
13. E. Schneider, A. New Type of Electrical Resonance, *Phil. Mag.*, **36**, p. 371, June, 1945.
14. E. Ginzton, Stabilized Negative Impedances, *Electronics*, **18**, pp. 140, 138, and 140 of July, Aug., and Sept., 1945, respectively.
15. J. Merrill, Theory of the Negative Impedance Converter, *B.S.T.J.*, **30**, p. 88, Jan., 1951.
16. H. Harris, Simplified Q. Multiplier, *Electronics*, **24**, p. 130, May, 1951.
17. J. Muchlner, Transfer Properties of Single and Coupled Circuit Stages With and Without Feedback, *Proc. I.R.E.*, **39**, p. 939, Aug., 1951.
18. F. B. Llewellyn, Some Fundamental Properties of Transmission Systems, *Proc. I.R.E.*, **40**, p. 271, March, 1952.
19. G. Fritzinger, Frequency Discrimination by Inverse Feedback, *Proc. I.R.E.*, **26**, p. 207, Feb., 1938.
20. R. Dietzold, Frequency Discriminative Electric Transducer, U. S. Patent 2,549,065, April 17, 1951.
21. R. Blackman, Effect of Feedback on Impedance, *B.S.T.J.*, **22**, p. 269, Oct., 1943.

Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath

By B. M. WOJCIECHOWSKI

(Manuscript received August 5, 1953)

A method has been recently developed for measuring thickness variations of a non-conductive cable sheathing, extruded over a grounded metal jacket. The method translates direct capacitance increments, sensed by probes sliding on the surface of the sheath, into thickness increments. The accuracy of the system based on this method is sufficiently high that the electrical error, which is of the order of a few thousandths of a μF , can be disregarded. Experimental data indicate that accuracy of the new system for absolute thickness measurements of homogeneous samples in stationary conditions is of the order of 0.002".

The error caused by translating capacitance to thickness depends on manufacturing elements and process tolerances, and can be evaluated on a statistical basis. Thus incremental measurements of the cable sheath thickness on the production line yield accuracies of the order of 0.003".

Application of this method to absolute sheath thickness measurements involves assumptions directly related to calibration and manufacturing process control. These aspects are rather extraneous to a measuring system per se, and, therefore, are not within the scope of this paper.

1 INTRODUCTION

1.1 The New Cable

A type of telephone cable has been developed in which lead is replaced with a polyethylene sheath extruded over a metal jacket. Since description of various aspects of this development can be found in the technical literature,^{1, 2, 3, 4} only some details of the cable construction and production that are pertinent to the understanding of the new measuring system, will be briefly outlined here.

The cable core, Fig. 1, is covered with a thin layer of a highly conductive metal, such as aluminum,¹ or two layers of different metals, such as aluminum and steel,^{2, 5} sealed longitudinally. To achieve the desired

mechanical properties, the metal jacket is corrugated circumferentially. Between the metal layer and the plastic sheathing, a bonding viscous thermoplastic compound is applied (Fig. 2). Normally, this compound fills the depressions of the corrugations on the metal surface adjacent to the surrounding polyethylene jacket.

The sheathed cable leaves the extruder with an essentially uniform speed, under pulling force of a capstan. For various sizes of cables and production settings, this speed may range from 30 to 80 feet per minute. After leaving the extruder, the cable is cooled in a trough of water and, before reaching the testing position, dried with compressed air.

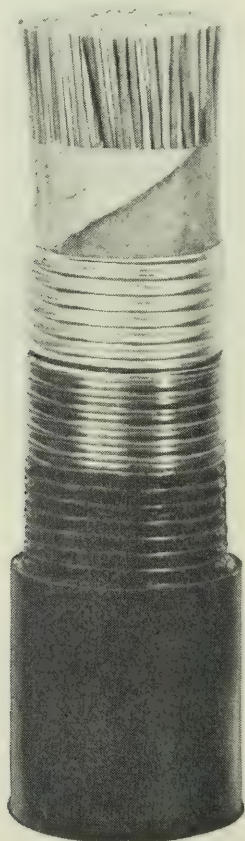


FIG. 1 — Polyethylene sheath telephone cable.

1.2 Measurement Difficulties

Some of the problems encountered in the manufacture of the new cable were related directly to the lack of reliable methods for measuring thickness of the plastic sheathing. Under manufacturing conditions, where sheath thickness cannot be adequately controlled, excess material must be used to assure meeting minimum thickness requirements.

Before the new method was developed, measurements were made by destructive testing of end samples. One or two circumferential strips were taken from each cable length and micrometer measurements were performed on each strip, at four to eight points. Unfortunately, the actual sheath thickness varies in a random way along the cable length, even between points only a few inches apart. It was evident that a method, based on a few point measurements, extrapolating long-cable properties which are describable rather in statistical terms only, left much to be desired.

1.3 Preliminary Considerations

The following methods of cable sheath measurements were considered:

- A. Use of an X-ray machine.
- B. Ultrasonic echo method (radar techniques).
- C. Capacitance measurements.

For practical reasons as well as for anticipated lack of accuracy, the first of these methods was rejected. The success of the second method was judged doubtful, the main reason being the presence of corrugations and of an irregular layer of the filling compound under the polyethylene sheathing, obscuring delimitation of the reflecting boundary surface. The third method, at first, also had discouraging aspects. In the case under discussion only grounded capacitance measurements are involved, since the metal core cannot possibly be insulated from the corrugating and forming machinery. The required long-time capacitance-to-ground stability and accuracy of the measuring system were estimated to be of the

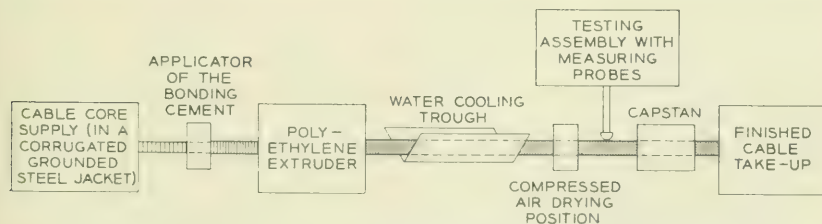


FIG. 2 — Block diagram of the polyethylene extruding process.

order of $0.001 \mu\mu\text{F}$ and $0.003 \mu\mu\text{F}$, respectively. Meeting requirements of this order, even under controlled laboratory conditions, presents some difficulties — and yet these requirements had to be met on a production line, on moving cable in the climatic and operational conditions prevailing in a large cable plant.

It was evident, therefore, that conventional grounded-capacitance measurements would not be practical. For instance, a shielded cable connecting the probes with the bridge circuit alone could produce wider random capacitance variations than the capacitance increments under measurement. Thus a new system which would meet all the necessary requirements had to be developed.

2 CIRCUIT DESCRIPTION

The measuring system which was developed consists of an impedance bridge, a phase sensitive detector, an unbalance indicator (recorder), capacitance probes and associated auxiliary equipment (See Figure 3).

2.1 *The Impedance Bridge for Grounded Direct Capacitance Measurements.*

The circuit shown on Fig. 4 employs a bridge having ratio arms⁶ magnetically coupled. An application of this type of circuit for capacitance measurements has been known for some time.⁷ Such a circuit is capable of performing in one balancing operation direct capacitance measurements while the center point (B) of the transformer ratio-arms winding is grounded. In our case, the "D" corner of the bridge consists of the metal covering of the cable core, which, as was mentioned above, is necessarily at the ground potential. Therefore, the "B" corner cannot be grounded. However, by connecting to this "B"-corner a shielding,⁸ surrounding the "A-D" and "C-D" measuring arms, including cables and probes, the following results can be achieved:

(a) Admittances from the measuring electrodes to the "B"-shielding are not critical. These admittances appear across the transformer-arms and, as a result of a close magnetic coupling realizable between these arms, any loading effects across any one of them are symmetrically reflected at the "A" and "C" corners of the bridge, thus essentially not affecting its balance.

(b) Stray admittances from the "B" shielding to ground appear across the opposite corners of the bridge (detector diagonal). Therefore, they also have no essential effects on the circuit balance.

(c) As a result of the "B"-shielding, stray admittances-to-ground

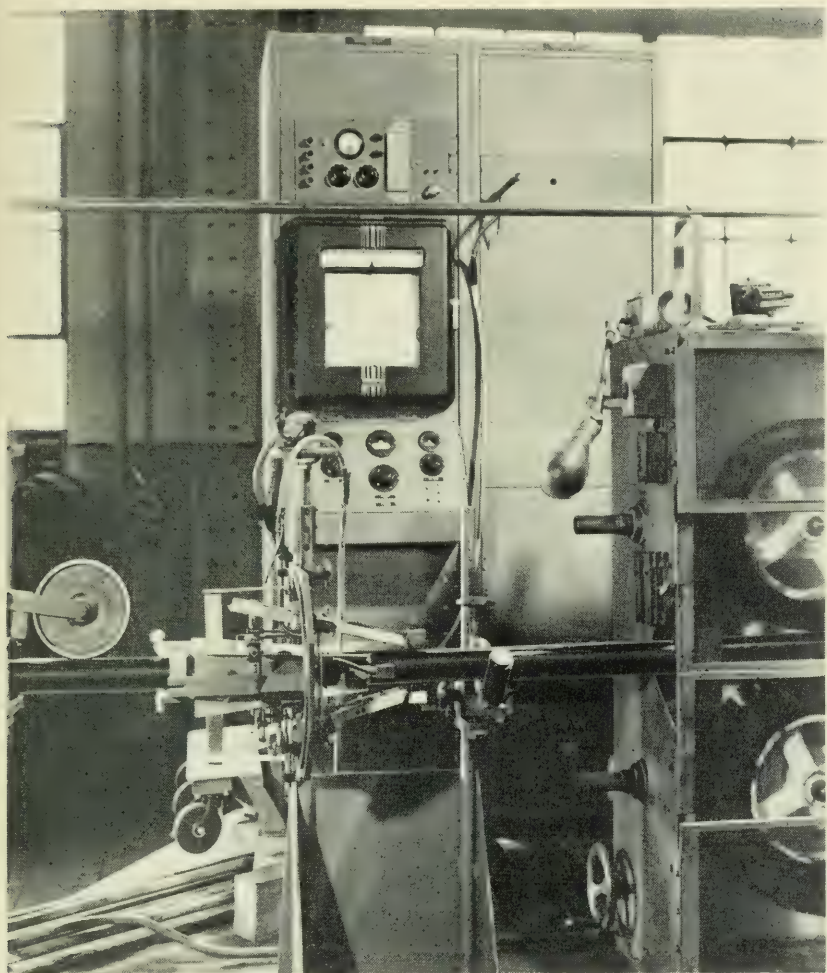


FIG. 3 — Measuring assembly.

from the measuring electrodes and from the connecting leads can be reduced to insignificant quantities.

As a result of the described circuit configuration, the bridge measures capacitance quantities equivalent to direct capacitance, in a particular case where one of two measuring electrodes is grounded. Realization of the grounded direct capacitance measurements is made possible by having within the measuring arrangement a three-electrode system in which stray admittances from the third (ungrounded) electrode to either

of the measuring electrodes do not affect the fundamental balance condition of the bridge network.

By the arrangement described not only are the residual effective capacitances between the measuring electrodes and ground reduced to a desirable minimum (actually below one $\mu\mu\text{F}$, including calibrating capacitor and balancing networks), but also any adverse capacitance effects of the cables connecting the bridge to the measuring probes are practically eliminated, even though these cables are several feet long.

The calibrated grounded direct capacitance range of the bridge extends over 0.32 $\mu\mu\text{F}$ in either direction off balance center position. Any unbalances within the $\pm 0.25 \mu\mu\text{F}$ range can be read in increments of 0.005 $\mu\mu\text{F}$ per division on a recorder. Since covering such a limited capacitance range directly by an adjustable capacitor could present various practical difficulties, a network, dividing electrically the range of a 100 $\mu\mu\text{F}$ differential capacitor by the ratio of 150 (approximately), has been applied. Using such a network facilitates calibration and adjustability and greatly reduces effects of the mechanical instability of the variable capacitor. (Similar networks are applied for capacitance and conductance residual balance controls.)

Stationary unbalances of the bridge network can be measured directly in a conventional manner by rebalancing the circuit with the calibrated capacitor. For unbalances rapidly varying in time, however, this null method could not be applied simply. Therefore, a proportional off-balance deflection method had to be used and various means to ascertain overall linearity between incremental capacitance unbalances and indicator deflections were provided, so that eventually variations in linearity no larger than 0.4 db over periods of several days and 0.2 db over several hours have been observed in the actual operating conditions.

Measurements with the bridge depend essentially on the calibrated capacitor. To avoid necessity for frequent and quite elaborate calibration checking (within a few one-thousandths of a $\mu\mu\text{F}$) of this capacitor in a laboratory, a set of supplementary, high stability auxiliary standards has been provided in the test set assembly. The capacitance values (1.05 $\mu\mu\text{F}$; 1.20 $\mu\mu\text{F}$; 1.35 $\mu\mu\text{F}$) of these capacitors are so chosen that differences between any pair of them can be compared directly with the calibrated capacitor in the bridge circuit. Reliability of this system is based on a reasonably high probability that change in the calibrated value of any single capacitor will be revealed in the process of mutually comparing all four capacitors. It was felt that this method of ascertaining calibration accuracy at the operating position was particularly recommended in the case of this circuit as its sensitivity to incremental capacitance unbalances

is actually higher than the sensitivity of the usually available laboratory equipment.

The bridge network is supplied by a 10-ke ac power source.

2.2 *Phase Sensitive Detector*

For eccentricity measurements and control of the sheathing process it is essential to register the direction of incremental deviations from an arbitrary level. For this purpose a phase sensitive detector^{9, 10} has been provided. Its simplified version is shown on Fig. 4.

By proper adjustment of the phase-shifter, the reactive component of the bridge unbalance signal can be oriented to be in-phase with the reference potential (b-a). In this condition, the capacitance unbalance sensitivity of the discriminator is at its maximum, and for a certain range of capacitance unbalances, linearity of the indicator may be assured. Also, when the above phase condition is fulfilled, the circuit is not sensitive to limited conductance unbalances (this fact also renders the circuit remarkably more stable than a similar circuit using a conventional null detector).

The dc output from the discriminator is fed through a balanced output stage (V2a and V2b), and an attenuator to a Leeds & Northrup zero-centered recorder. At the operating sensitivity level, each of the 100 divisions of the recorder scale corresponds to $0.005 \mu\mu F$, or approximately to 0.001 inch of the incremental sheath thickness. The rôle of the attenuator is two-fold: it provides control of the over-all sensitivity of the measurements (in steps of 0.2 db), and it introduces more than 20 db attenuation into the dc output signal path. This loss is compensated by an added gain within the feedback-controlled ac amplifier (AC-A) preceding the phase discriminator. The net result of this "ac for dc gain-trading" is a considerable improvement of the over-all circuit stability since the range of random drifts, such as usually generated within the phase-discriminator and its direct-coupled output stage, are materially reduced.

2.3 *Measuring Probes*

As has been mentioned above, two arms of the bridge circuit consist of a pair of admittances between the grounded metal core of the cable (D corner) and the probes sliding on the surface of the plastic cable sheathing. These probes are connected to the "A" and "C" corners of the bridge, respectively, with two shielded flexible conductors (each

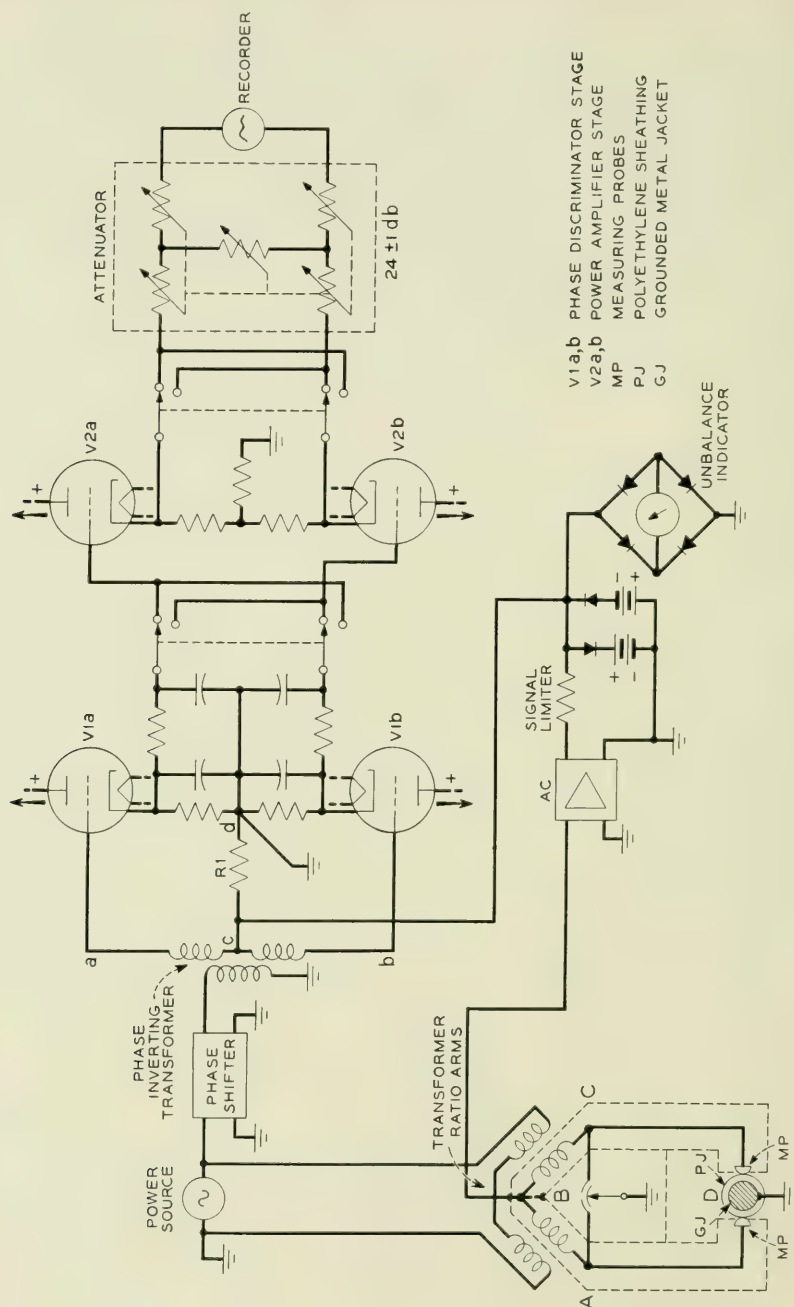


FIG. 4 — Bridge and discriminator circuit.

about 10 feet long) and are maintained mechanically in the testing position by the probe assembly (see Figs. 3 and 5(a)).

In the design of the probes and their assembly, various difficulties had to be overcome. The probes operate on cables subjected to some unavoidable swings and vibrations while moving with speeds up to 80 feet per minute. The capacitance from either of these probes to the metal cable core, in equivalent conditions, should match each other within approximately one-thousandth of a $\mu\mu\text{F}$. This capacitance should not be appreciably affected by limited displacements of the probes with respect to the cable plane of symmetry, such as may occur in actual operating conditions.

The first experiments with probes of a conventional design, having flat, or nearly flat, contact surfaces, were quite discouraging. The probe-to-core capacitances fluctuated to an intolerable degree as a result of even minute cable displacements.

Eventually, probes were developed which met all the requirements. Each of these probes is in the form of a cut-off segment of a toroid. The major axis of the cut-off elliptical plane is oriented in the direction essentially parallel to the cable axis, while the convex center part of the probe slides on the cable sheathing. This form of probe has the advantage, common with the spherical form, that the capacitance from the probe to the cable core varies but little as a result of displacements and changes of position caused by the cable motion. But the toroidal form has the following advantages over the spherical: first, for the same residual capacitance to the cylindrical cable core, the transverse dimensions of the former are smaller; and, second, the capacitance of the toroidal form with respect to a cylindrical cable core can be conveniently adjusted by the simple expedient of twisting the probe element in a plane parallel to the cable axis. (Adjustments with a precision exceeding one-thousandth of a $\mu\mu\text{F}$ were actually performed).

The probe electrodes, surrounded (except for the contacting face) by the B-shielding, are mounted on mechanically balanced light aluminum arms [Figs. 5(a) and 5(b)]. There might be one, two, or four probes to an assembly, which can be turned over 360° around the cable axis. For eccentricity measurements two probes can be simultaneously used, having a spacing of 180° (for measurement of eccentricity across a diameter) or of 90° (for measurement of ellipsoidal eccentricity). Also for eccentricity or direct thickness investigations and process settings one probe only may be used, with the other bridge measuring arm connected to an auxiliary standard.

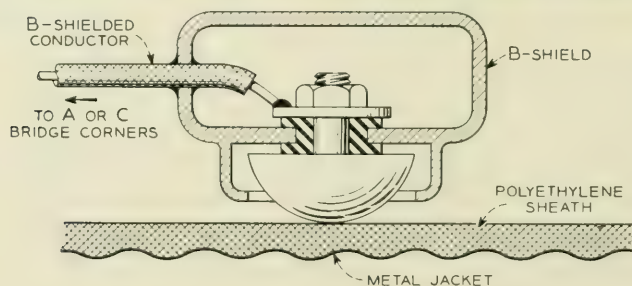
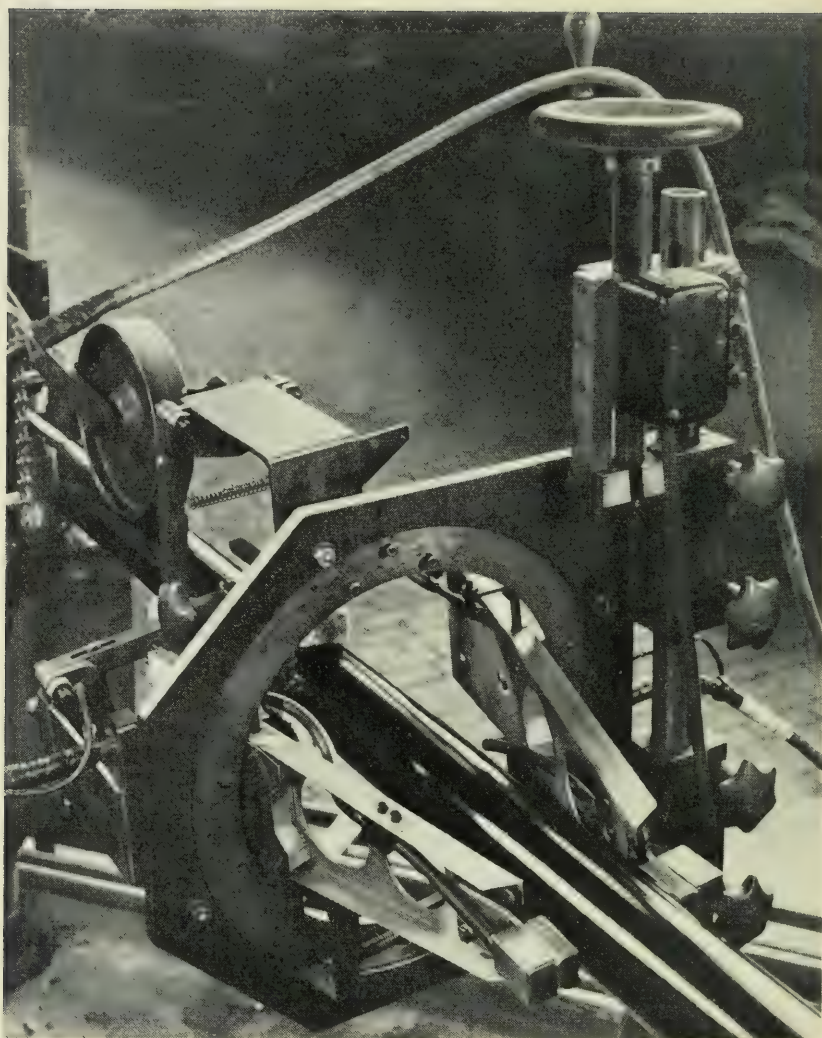


FIG. 5 — (a) Measuring probe assembly. (b) Probe element.

The average capacitance from the probe element to the grounded metal core varies from 1.1 to 1.3 $\mu\mu\text{F}$ for cables measured.

3 EXPERIMENTAL RESULTS

3.1 Circuit Performance Under Stationary Conditions

Incremental capacitance sensitivity for grounded direct capacitance measurements, in normal operating conditions with the probes in contact with a cable sample: order of 0.001 $\mu\mu\text{F}$.

Circuit stability and repeatability for periods over one hour duration: $\pm 0.003 \mu\mu\text{F}$.

Overall linearity of the unbalance indications, as read on the recorder scale within the range of plus or minus 0.25 $\mu\mu\text{F}$ off center-balance position: $\pm$ (3 per cent + 0.003 $\mu\mu\text{F}$).

Mechanical Stability: Moving or twisting of the connecting leads has no effect on balance stability. Swinging of the cable under measurement, even beyond the limits encountered in actual working conditions, produces barely noticeable effects on the balance indication.

Capacitance measurements on flat polyethylene samples: One of the measuring bridge arms was connected to the auxiliary standard of 1.20 $\mu\mu\text{F}$. The other arm was terminated by the probe in contact with a flat polyethylene sample placed on a grounded metal-plate. Thickness of samples at the point of contact was measured with a micrometer to the nearest 0.0005 inch. Capacitance unbalance readings were taken directly on the recorder scale to the nearest 0.005 $\mu\mu\text{F}$. In order to avoid noticeable "air-gap" and "surface" effects, which occur when stacking several samples, in no case were more than two flat samples in a stack measured. Under these conditions, repeatability of readings was within one recorder division (0.005 $\mu\mu\text{F}$), equivalent approximately to one-thousandth of an inch. In a typical case shown on Fig. 6, of 38 measurements taken in the thickness range from 0.052 inch to 0.168 inch, only three measurements were off from the averaging curve by more than 0.002 inch. (Further investigation disclosed that these three points, marked " Δ " on Fig. 6, were all associated with a particular sample.)

Capacitance measurements on stationary cable samples. In order to establish statistical reliability of measurements on actual cables by the described capacitance method, a number of cable samples were tested, varying in core diameter, average polyethylene sheathing thickness and mechanical construction.

A typical graph resulting from plotting capacitance increments versus micrometer measurements of a cable sample is shown on Fig. 7. Out of

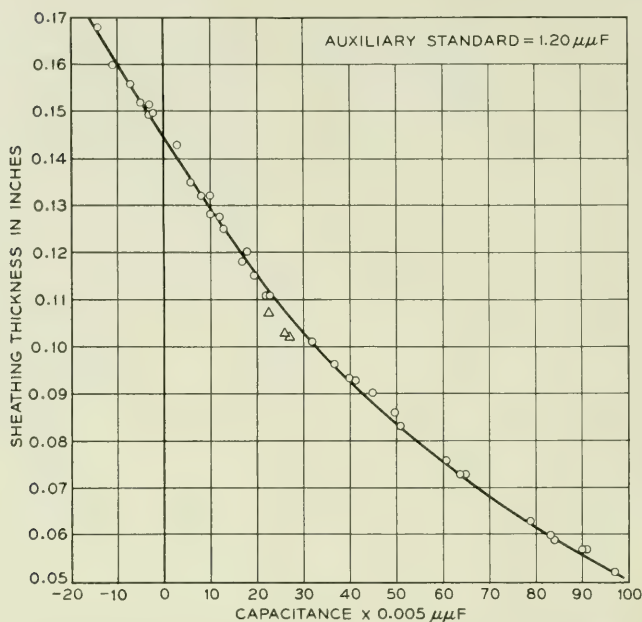


FIG. 6 — Capacitance versus thickness measurements of polyethylene plate samples.

25 measured points, 21 are contained within ± 0.003 inch limit off an average curve. Three out of four remaining points (marked "Δ") were found to be from cable areas where application of the thermoplastic cement was excessive (explanation for an extraneous position of the fourth point had not been found by the time these measurements were concluded).

On the basis of over 800 measurements it has been estimated that at least 75 per cent of the plotted points are within the limits of ± 0.003 inch deviation from an average capacitance versus thickness curve, for samples taken from the same cable. For samples taken from different cables these deviations ranged sometimes up to ± 0.005 inch. A few points (less than 5 per cent) showed deviations larger than 0.01 inch. With some rare exceptions, these extreme deviations indicated larger than actual sheath thicknesses, and in most of the cases they were associated with areas where an excess of the flooding cement was present.

3.2 Incremental Capacitance Measurements on the Production Line

Experimental measurements on over 500 feet of cable length were performed on the production line with one of the bridge arms connected

to the probe, and the other arm to an auxiliary standard of nominal value $1.20 \mu\mu\text{F}$. The probe was placed on the surface of a cable moving with a speed of approximately 50 feet per minute. The line, along which the probe was sliding over the sheath, was marked for subsequent measuring purposes. At discrete intervals, the angular position of the probe with respect to the cable circumference was advanced by an angle of 90° . The unbalance signals were traced on the recorder chart with a standard sensitivity of $0.005 \mu\mu\text{F}$ per division. After completion of the cable run, the sheath was stripped and washed (to remove flooding compound) and micrometer measurements were taken along the probe route at points six inches apart. Subsequently these micrometer measurements were plotted in scales equivalent to the recorder chart.

A typical example of a measurement performed on a cable section approximately 250 feet long (a total of 500 measured points) is shown on Fig. 8. The upper curve represents a photograph of the recorder tracing. The lower curve was obtained by connecting point-to-point actual thickness readings and plotting them on the non-linear vertical scale following the capacitance versus thickness function (similar to that as shown on Fig. 6), to make both charts graphically equivalent.

From comparison of these graphs a few observations can be made. In fact, these curves represent fundamentally different methods of derivation. The recorder indications are continuous average readings based on an area having a definite width and a length of a few corrugation spaces

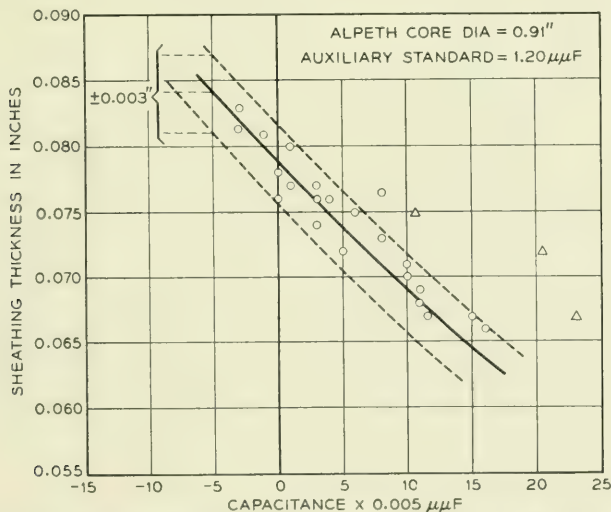


FIG. 7 — Capacitance versus thickness measurements of cable sheathing sample.

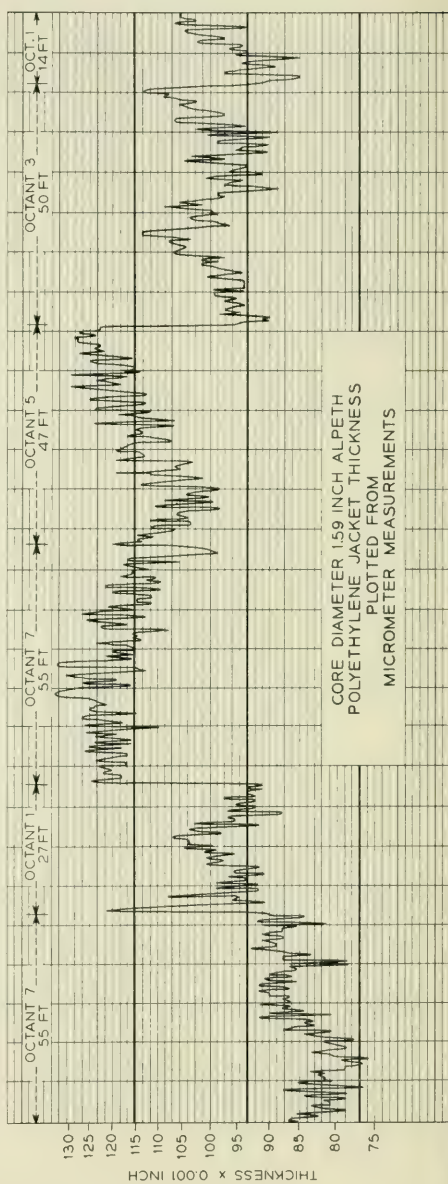
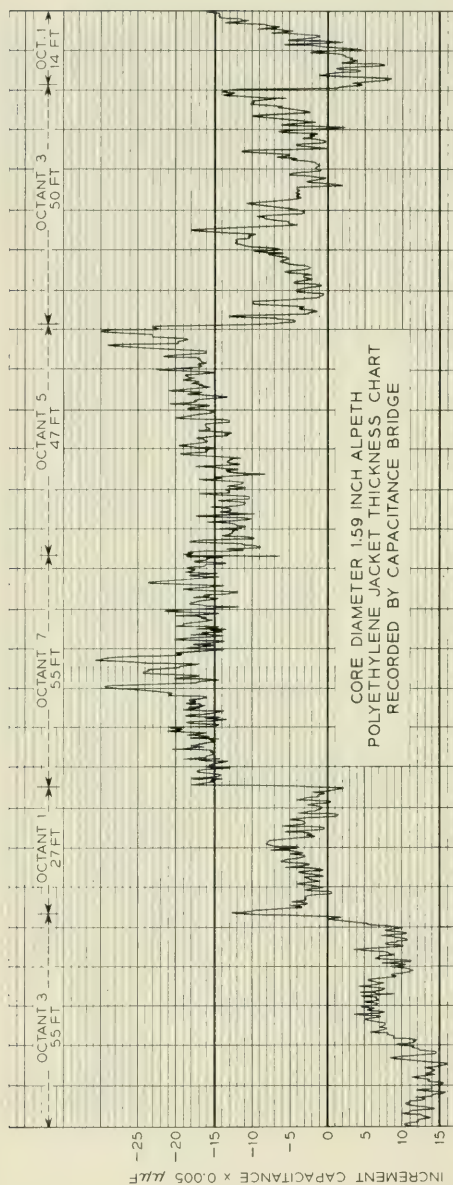


FIG. 1. Core diameter 1.59 inch alpeth polyethylene jacket thickness chart plotted from micrometer measurements (upper curve) with the micrometer measure-

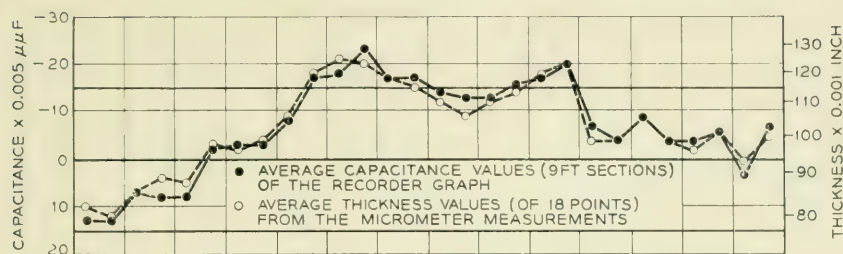


FIG. 9 — Comparison of average values from graphs of Fig. 8.

while the micrometer readings are point measurements taken at discrete distances at the bottom of the corrugation valleys in the polyethylene jacket. Despite this fact, the statistical character of both graphical results is closely similar (See Fig. 9). Assuming an average translation factor of $0.005 \mu\mu\text{F}$ per 0.001 inch, and discarding tracing errors, the agreement for incremental measurements between both methods can be estimated to be of the order of 0.003 inch. This accuracy is ample for any practical purpose of incremental thickness control of cable sheathing.

4 CONCLUSIONS

The method presented here for non-conductive sheath thickness measurements yields sufficient stability and translation reliability to be considered an improvement in the art. In particular, the incremental capacitance measurement accuracy of the order of $0.003 \mu\mu\text{F}$ (equivalent to less than 0.001 of an inch of incremental sheath thickness) is sufficiently high to disregard, in practical applications, the error of the test set itself. When measuring flat samples in stationary conditions accuracies of the order of 0.002" for absolute thickness are obtainable.

Reliable differential measurements of cable sheathing on the production line can be realized on a statistical basis. Accuracies of the order of 0.003 of an inch for incremental sheath thickness (eccentricity) measurements were consistently attained in actual manufacturing conditions.

Extensive experience with the described measuring system indicates that it can also be applied for absolute sheath thickness measurements yielding desirable accuracies. This application, however, involves various manufacturing and process control problems extraneous to the measuring system per se. These aspects, therefore, are not discussed in the present paper.*

* Development of techniques for absolute sheath thickness measurements, using the described system, is being conducted by W. T. Eppler of Western Electric Co., Inc., Kearny, N. J.

5 ACKNOWLEDGMENTS

In carrying the development work described in this paper, I was helped considerably by fellow engineers as well as men at various supervisory levels of the Western Electric Company. In particular, I would like to thank R. I. Neel for assistance in gathering experimental data at the initial stages of this work, W. T. Eppler for his active part in confronting operation of the system with production line conditions, and D. T. Robb for encouragement and help given me during the development work and in preparation of this paper. A special credit for providing a mass of laborious statistical data concerning accuracy of the method in actual manufacturing conditions belongs to J. L. O'Toole of Bell Telephone Laboratories' group at Kearny.

BIBLIOGRAPHY

1. R. P. Ashbaugh, Alpeth Cable Sheath, Bell Lab. Record, **26**, pp. 441-444, Nov., 1948.
2. R. P. Ashbaugh, Stalpeth Cable Sheath, Bell Lab. Record, **29**, pp. 353-355, Aug., 1951.
3. F. W. Horn and R. B. Ramsey, Bell System Cable Sheath Problems and Designs, Trans. A.I.E.E., **70**, Pt. 1, pp. 1811-1816, 1951.
4. V. T. Wallder, Polyethylene for Wire and Cable, Elec. Eng., **71**, pp. 59-64, Jan. 1952.
5. U. S. Pat. 2,589,700, H. G. Johnstone (1952).
6. J. G. Ferguson, Classification of Bridge Methods of Measuring Impedances, B.S.T.J., **12**, pp. 452-468, Oct. 1933.
7. C. H. Young, Measuring Inter-Electrode Capacitances, Bell Lab. Record, **24**, pp. 433-438, Dec., 1946.
8. L. Hartshorn, *Radio-Frequency Measurements by Bridge and Resonance Methods*, Chapter III, John Wiley & Sons, New York, 1940.
9. A. C. Seletzky, Cross Potential of a 4-Arm Network, Elec. Eng., pp. 861-867, Dec., 1933.
10. U. S. Pat. 2,554,164, B. M. Wojciechowski (1951).

The Application of Designed Experiments to the Card Translator

By C. B. BROWN and M. E. TERRY

(Manuscript received October 27, ¹1953)

In the course of development of the card translator for use in the No. 4A toll crossbar system it was necessary to evaluate in detail the performance characteristics of all phases of the translator operation. One of the most important phases was that of the action of the translator cards. Since the action of these cards is controlled by the simultaneous influence of many independent variables a study of the card action was made using statistically designed experiments. This study made use of Graeco-Latin Square and Factorial Designs; herewith is presented a detailed description of their application to the problem. Also the method of analysis of the data and resulting conclusions are described in detail.

I. INTRODUCTION

In order to permit circuit designers to use the full capabilities of card translators¹ it has been necessary to conduct tests to determine (1) the time intervals required to drop the cards into reading positions and (2) the maximum number of cards which a translator can operate reliably. It was also desired to establish the maximum and minimum number of cards that could be operated efficiently in a card translator. Early estimates indicated a machine capacity of 1,000 cards but if a higher capacity could be demonstrated, a considerable cost saving in the 4A system would result. Since some differences in card drop time known to exist are due to card position and machine loading this study also was to consider whether or not any special loading instructions were necessary or desirable for field use of card translators. As the study commenced the Western Electric Company Installation Department requested information as to the need for leveling card translators on installation and it was felt that such information could readily be obtained in the course of this study. In investigating the effect of the many variables related to card drop time, a consideration was given to the possibility of the need for any design changes.

All previous tests of card drop time were made on Laboratories' built models of the translators and since tests made after any changes in model design have produced considerable differences in card drop times, it was felt that this study should be made on a sample of the Western Electric product. In addition, previous tests did not encompass all of the known variables over their extreme ranges.

For this study, a new translator was obtained from Western Electric. This translator was one of the regular production and was selected to be a representative unit. Fourteen hundred new 200A blanks were also obtained for use in this translator. Of these, 10 were coded for use as test cards to be observed. One hundred were coded to fill some bins with all coded cards. The balance of the 1,400 cards was left uncoded. This translator was connected to a simplified test set which would cycle the machine through the operation of the 10 coded test cards. Although this test set was simplified in its operation all of the pertinent time relationships involving card dropping were the same as are used in the standard translator circuits.

This study considered all of the variables, with the machine working in a normal cycle of operation as is currently used in the 4A system. These variables are:

1. *Bin* — Some differences in card dropping time had been observed depending on the bin in which the operating card was located.

2. *Position of Card Within Bin* — Earlier tests indicated that the location of the card within its bin made a considerable effect on dropping time.

3. *Code of Card* — Since both three and six digit cards will be used in the translator the code becomes a variable which may affect dropping time.

4. *Coded or Uncoded Cards in Working Bin* — Previous tests were made using only a few coded cards in the bin under observation and it was felt necessary to learn whether or not having all coded cards in a bin made any difference in the dropping time.

5. *Load in Working Bin* — This relates to the number of cards in the bin under observation. A range of from 15 to 105 cards per bin was used.

6. *Load in Machine* — This has to do with the possibility of any effect into the bin under observation from the cards in the other bins of the translator.

7. *Consistency of Data* — This has to do with repeated measurements to observe if the card dropping time is consistent over short periods of time.

8. *Balanced versus Unbalanced Loading*—Since the earlier installations will have machines that are less than full, it was necessary to learn whether or not any special loading considerations with regard to the position of the cards should be specified.

9. *Tilt*—(Machine) This has to do with the accuracy with which the translators should be leveled on installation.

10. *Life*—This variable considers the effect of repeated operations on cards and what effect, if any, these repeated operations have on the dropping time.

The tests on the first nine of the variables have been completed and a discussion of their effect on card dropping time is possible at this time. The tests on the tenth variable, life, are still in progress and have not as yet proceeded far enough to draw any conclusions. Therefore no consideration will be given to the effects of usage in this article. The effect of most of these variables was considered when the translator was operated both with and without the card support bars being used. The card support bars operate at the same time the card is dropped. Their operation is required for every card in the machine. Since their operation is slower than the free fall time of the operating card the card may ride down into the operated position on these bars and the observed drop time will simply be the operate time of the card support bars. This masks the effect of the various friction, gravity, and magnetic forces on the cards. Operating without the card support bars gives information on the free dropping time of the cards uninfluenced by these bars. Thus the masking effect of the card support bars can be removed and the tests made more sensitive to any physical effects which may be taking place at the cards themselves.

II. DESIGN OF EXPERIMENTS

Theory of Designed Experiments

In this problem it was required to evaluate the effect of each of the nine variables independently. It might have been convenient to take one variable at a time and vary it over its range while holding the associated variables at sets of constant values. Then a second independent variable could be chosen, the first variable placed in the group of associated variables, and a new set of runs made. This could be repeated until all the variables had been tested. This is obviously a logical, straightforward but unwieldy procedure.

A simpler procedure would be to vary the chosen variable over its range and at the same time let the associated variables vary over their

assigned ranges in such a way that each associated variable is set an equal number of times at each of its possible values. This will cause any effect of the associated variable to be balanced out with respect to the chosen variable. That is, the average response of the system to the chosen variable can be evaluated because the contributions to the response from the associated variables are balanced out with respect to their variations.

Mathematically stated, the purpose of the latter procedure is to allow the independent evaluation of each variable over the observed range of its variation in the presence of the other variables.

Until recent years the standard experimental procedure was to make measurements for several values of one variable while taking great precautions to hold everything else constant. It was early recognized that such a technique was expensive but until the recent development of experimental designs in the field of Statistics there was no known alternative. In the class of designs treated here,² the set of numbers representing the respective sums of the measurements taken at the selected values of a given variable has the following properties:

1. If there are K discrete values of the given variables and N measurements made, then each sum contains N/K measurements.

2. If there are K_i values of the i^{th} remaining variable then N/K must be divisible by K_i , for all i .

3. Within each sum, for any given variable, each of the discrete values of the remaining variables must occur the same number of times.

4. The set of values of any of the remaining variables occurring in any sum for a given variable must be identical.

These properties imply that the logical subgroups for the discrete values of any given variable all contain the same number of measurements balanced with respect to the values of all the remaining variables.

The joint simultaneous evaluation of the specified variables can thus be made since it can be shown that the logical subgroups with respect to a given variable are independent of the logical subgroups of any other variable.

One of the difficulties in the use of this new technique is that both an engineer and statistician are required, and only rarely are both these professions found in an individual. It therefore becomes the responsibility of the engineer to outline the major and minor variables of the experiment, the description of the measuring devices to be employed, the accuracies desired in the results and his budget. The statistician must then propose the types of designs that will be appropriate together with their costs, methods of analysis, and salient features. The engineer and statistician must then select the design which best fits the situation.

The design selected for the first experiment of this study combined two basic design types, the Latin Square,² and the Factorial Design.² A brief discussion of each is in order at this point.

The Latin Square is peculiarly well suited to engineering research problems where many variables exist but the number of discrete levels necessary to describe the variation of any one is small, where relatively great precision is possible in measurements, and where the interaction of the variables is not a factor of the experiment. A key design² for a 5x5 Latin Square is as follows:

	Column				
Row	1	2	3	4	5
1	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
2	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>A</i>
3	<i>C</i>	<i>D</i>	<i>E</i>	<i>A</i>	<i>B</i>
4	<i>D</i>	<i>E</i>	<i>A</i>	<i>B</i>	<i>C</i>
5	<i>E</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>

It will be observed that each Latin letter falls once and only once in every row and column. It is also true that if the logical subgroups *A*, *B*, *C*, *D*, *E* are considered, each of these sums has each row and column included once. If to the five row values, column values and letters, the five values of the first, second and third variables are associated respectively, the variables represented by Row, Column, and Letter can be evaluated independently of each other.

On the Key Latin Square, another properly chosen square represented by Greek Letters can be superimposed, and the five values of the fourth variable assigned to these letters at random:

	1	2	3	4	5
1	α	β	γ	δ	ϵ
2	γ	δ	ϵ	α	β
3	ϵ	α	β	γ	δ
4	β	γ	δ	ϵ	α
5	δ	ϵ	α	β	γ

The Graeco-Latin Square below is then produced:

	1	2	3	4	5
1	<i>A</i> α	<i>B</i> β	<i>C</i> γ	<i>D</i> δ	<i>E</i> ϵ
2	<i>B</i> γ	<i>C</i> δ	<i>D</i> ϵ	<i>E</i> α	<i>A</i> β
3	<i>C</i> ϵ	<i>D</i> α	<i>E</i> β	<i>A</i> γ	<i>B</i> δ
4	<i>D</i> β	<i>E</i> γ	<i>A</i> δ	<i>B</i> ϵ	<i>C</i> α
5	<i>E</i> δ	<i>A</i> ϵ	<i>B</i> α	<i>C</i> β	<i>D</i> γ

Note that now each of the variables associated with Row, Column, Latin letter, Greek letter has the property that each element of the four categories contains one and only one element of the remaining three categories. For example, consider the five Latin-Greek letter combinations, or sample cells, containing ϵ :

Row 1 Col 5, *E*; Row 2 Col 3, *D*; Row 3 Col 1, *C*; Row 4 Col 4, *B*; Row 5 Col 2, *A*.

Then the sum of these five cells will contain the contribution of the 5 Rows, 5 Columns, and 5 Latin letters.

It should be noted that the rows and columns can be permuted without affecting the properties of the square. Indeed to protect against systematic effects which may be detrimental, it is usual to assign at random the row and column number, as well as the Latin and Greek letters to the values of variables represented by them. A notable exception to randomization occurs when time is a variable and those measurements made under essentially the same conditions within the same unit of time become the experimental unit. In the first experiment, all measurements made on one Run become the experimental unit with respect to time.

The Factorial Design serves a different purpose. In this discussion only two independent variables X , Z will be predicated but the extension to more variables follows directly. The XZ plane is the plane of the independent variables and we seek the point (X_0, Z_0) which gives a value of y (the dependent variable), $Y = \varphi(X, Z)$, which is optimum in some sense. That is, $\min Y$, $\max Y$ may be sought, or the surface $f(X, Y, Z)$ shown to be a plane.

Generally only the region in the neighborhood of y (optimum) is of interest to the experimenter. Hence it is imperative (1) to bracket this point with respect to each independent variable and (2) to have a method of estimating y (optimum). If a factorial experiment has l_x different values of X and l_z different values of Z , then each replication of the experiment will require $l_x \cdot l_z$ units or points (X, Z) . The first repetition of an experiment is called the second replication in the same way that the first overtone in music is called the second harmonic by engineers. Since the number of units available for test is usually limited, this places a practical ceiling on the magnitude of l_x and l_z . As a practical limit in general l should be 7 or less and the values 2, 3, or 4 are far more common. It is generally better to use the smaller values of l and repeat the experiment, than to conduct an experiment involving only a single replication. In addition to evaluating one variable averaged over the second, we

are interested in evaluating the interaction of the variables on each other, when such interaction exists and is of interest. In a sense this interaction measures the departure of the system y , X , Z from linearity.

Once the basic designs have been selected and appropriately combined to fit most efficiently the requirements of the proposed experiment, and the values of the variables randomly assigned to the schematic layout, a detailed experimental layout must be drawn up. This layout must show concisely and clearly each experimental unit and the makeup of every basic element giving its assigned value of each variable. Explicit directions must be drawn up as to the order of selection of the elements of the unit. It is generally advisable for simplicity to assign the elements at random to the M possible consecutive order integers of the experimental unit.

Performance Study

The first seven variables listed in the Introduction are:

1. The Bin in use, *Bins*.
2. Position of Test card pair within the bin, *Position*.
3. The use of 3 digit or 6 digit Cards, *Code*.
4. Arrangement of Coded and Uncoded Cards, *Runs*.
5. Load of Bins containing Test Cards, *Load*.
6. Load of Bins not containing Test Cards, *Idlers*.
7. Order of repeated measurement, *Look*.

These constitute a system — that is, any or all can be varied at will and hence a design involving all of them simultaneously can be sought.

The test set can operate ten test cards; 5 with a 3 digit code and 5 with a 6 digit code. This immediately suggests 5 packages of two coded card pairs, each pair containing a 3 and a 6 digit card. Five pairs can also be handled neatly in 5 bins. The combination of the standard load of 85 cards, with two overloads and two underloads would give a fair evaluation of load criterion. Budget restrictions force the use of only a limited number of coded cards, with blanks used to fill out the experiment. Hence the type of card making up the load must be varied over the loads. It was further found that 5 positions of test cards within the bin covers the range of positions adequately.

The pairs of 3 and 6 digit cards now are associated with the Graeco-Latin Square design with Columns identified with Bins; Rows with distribution of coded cards; or Runs; Latin letters with Load; and Greek letters with position within the bin. Now if the position of the 3 digit card is randomly assigned in the pairs, the design absorbs the first five

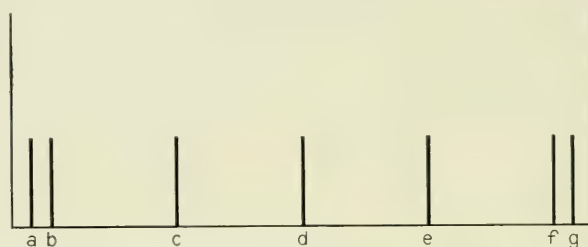
TABLE I—DESIGN OF GRAECO-LATIN SQUARE EXPERIMENT

Each run made with the "x" bins loaded with 0, 50, 85 and 100 cards and consists of four operations (A, B, C, D) of each coded card.

Bin No.

Photocell Mounting	I	II	III	X	IV	V	X	X	X	X	X	X	Lamp
-----------------------	---	----	-----	---	----	---	---	---	---	---	---	---	------

POSITION IN BIN



		Bin I	Bin II	Bin III	Bin IV	Bin V
Card →		1 6	2 7	3 8	4 9	5 10
Run #1	Pos. in Bin.....	<i>d g</i>	<i>f g</i>	<i>a b</i>	<i>a d</i>	<i>c e</i>
	No. of Coded Cards...	2	41	50	15	2
	No. of Blank Cards...	98	44	0	0	103
Run #2	Pos. in Bin.....	<i>c e</i>	<i>d g</i>	<i>f g</i>	<i>a b</i>	<i>a d</i>
	No. of Coded Cards...	85	9	15	2	2
	No. of Blank Cards...	0	41	0	103	98
Run #3	Pos. in Bin.....	<i>f g</i>	<i>a b</i>	<i>a d</i>	<i>c e</i>	<i>d g</i>
	No. of Coded Cards...	2	100	2	2	7
	No. of Blank Cards...	103	0	83	48	8
Run #4	Pos. in Bin.....	<i>a d</i>	<i>c e</i>	<i>d g</i>	<i>f g</i>	<i>a b</i>
	No. of Coded Cards...	2	2	105	2	2
	No. of Blank Cards...	48	13	0	98	83
Run #5	Pos. in Bin.....	<i>a b</i>	<i>a d</i>	<i>c e</i>	<i>d g</i>	<i>f g</i>
	No. of Coded Cards...	2	2	2	85	22
	No. of Blank Cards...	13	103	98	0	28

variables of the list. The layout of these variables in the Graeco-Latin Square design is given in Table I.

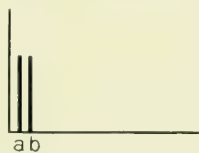
The remaining two variables are functions of different portions of the machine. The rows of the Square, involving distribution of coded and uncoded cards, represent distinct machine set-ups, and hence if for every set-up the load of the seven not-measured bins is varied, and for every variation in this load 4 observations of the 5 pairs are taken,

consideration of the remaining two variables is achieved. Yet the tedious part of handling the test cards has been reduced to a very reasonable amount. These last two variables, considered by themselves, form a factorial design and where 4 loads and 4 measurements per load are used, 16 measurements being taken for each machine set-up. Similarly any of the other five variables considered pairwise with one of the latter two forms another factorial design. This resulting complex overall pattern is shown in Table I.

TABLE II — TRANSLATOR TILT TEST — DESIGN OF EXPERIMENT
TEST CARDS LOCATION

Bin	I	II	III	IV	V
Cards	6, 1	2, 7	8, 3	4, 9	10, 5

POSITION OF TEST CARDS IN BINS



Total number of cards in each bin: 100

Bin # I contains all coded cards.

	Description of Tilt
Tilt #0	Lamp end lower than photocell end by $\frac{1}{16}$ " in 3 feet. Translator resting on table at all four points. North Lamp end higher than South Lamp end by $\frac{1}{16}$ " in 22". North Photocell end higher than South Photocell end by $\frac{1}{16}$ " in 22"
Tilt #1	Lamp end higher than photocell end by $\frac{9}{16}$ " in 3" ft. Lamp end supports blocked up $\frac{1}{4}$ " North Lamp end higher than South Lamp end by $\frac{1}{8}$ " in 22" North Photocell end higher than South Photocell end by $\frac{1}{8}$ " in 22"
Tilt #2	Lamp end higher than photocell end by $1\frac{5}{16}$ " in 3 ft. Lamp end supports blocked up $\frac{1}{2}$ " North Lamp end higher than South Lamp end by $\frac{3}{16}$ " in 22" North Photocell end higher than South Photocell end by $\frac{3}{16}$ " in 22"
Tilt #3	Lamp end higher than photocell end by $1\frac{1}{8}$ " in 3 ft. Lamp end supports blocked up $\frac{5}{8}$ " North Lamp end higher than South Lamp end by $\frac{3}{16}$ " in 22" North Photocell end higher than South Photocell end by $\frac{3}{16}$ " in 22"

Tilt Study

Routine field practice for the installation of a card translator calls for leveling the table before positioning of the translator. After a translator was placed it had been found that the level was not maintained, and the question of final level requirements was raised. Accordingly the test machine was run at 0 in., $\frac{5}{8}$ in., 1 in., and $1\frac{3}{16}$ in. tilt, Table II. Two test cards were placed in each of five bins. Since the two end card positions on the low side of each bin were suspected as being critical, the two test cards were placed in these positions in the five bins. The experiment is shown schematically below:

	Bins				
	I	II	III	IV	V
Tilt 0	a_6b_1	a_2b_7	a_8b_3	a_4b_9	$a_{10}b_5$
1	a_6b_1	a_2b_7	a_8b_3	a_4b_9	$a_{10}b_5$
2	a_6b_1	a_2b_7	a_8b_3	a_4b_9	$a_{10}b_5$
3	a_6b_1	a_2b_7	a_8b_3	a_4b_9	$a_{10}b_5$

where a is the end position, b is the next-but-end position, and the subscripts identify the actual card number. Considering any bin, the two card positions and the four values of tilt may be considered as a factorial design. Similarly, the five bins combine with the four values of tilt as a factorial design.

Balanced Loading

At the onset of general usage many translators will not be fully loaded. It was desirable to investigate the effects of various patterns of loading at three representative low loads of 200, 400 and 600 cards respectively. Four logical patterns were studied (see Tables III and IV) all of which are shown below for a load of approximately 400 cards. (Similar patterns follow with loads of 200 and 600 cards.):

Bin	1	2	3	4	5	6	7	8	9	10	11	12
W	x	x	x	x								
X	x	x					x	x				
Y	x	x									x	x
Z	0	0	0	0	0	0	0	0	0	0	0	0

$x = 100$ cards, $0 = \frac{1}{2}$ load (approximately)

The ten test cards were all placed in Bin One.

This experiment involves only two factors; the four loading patterns,

TABLE III — LOADING PATTERNS IN PARTIALLY LOADED
CARD TRANSLATOR — BALANCED VERSUS UNBALANCED
LOAD TEST

Treatments	Load	No. of Cards in Bins											
		Bin I	Bin II	Bin III	Bin IV	Bin V	Bin VI	Bin VII	Bin VIII	Bin IX	Bin X	Bin XI	Bin XII
W	600	100	100	100	100	100	100	—	—	—	—	—	—
	400	100	100	100	100	—	—	—	—	—	—	—	—
	200	100	100	—	—	—	—	—	—	—	—	—	—
X	600	100	100	100	—	—	—	100	100	100	—	—	—
	400	100	100	—	—	—	—	100	100	—	—	—	—
	200	100	—	—	—	—	—	100	—	—	—	—	—
Y	600	100	100	100	—	—	—	—	—	—	100	100	100
	400	100	100	—	—	—	—	—	—	—	—	100	100
	200	100	—	—	—	—	—	—	—	—	—	—	100
Z	600	50	50	50	50	50	50	50	50	50	50	50	50
	400	33	33	33	33	33	33	33	33	33	33	33	37
	200	16	16	16	16	16	16	16	16	16	16	16	24

Note: All cards in Bin I are coded, see Table IV. Cards in Bins II-XII not coded.

and the three loads of cards. A suitable design is the factorial design with one factor, the loading pattern, at four levels (W, X, Y, Z) and the other factor, the loads of cards at three levels (200, 400, 600) with ten test card measurements taken at each of the twelve points.

III. DATA

The data on card dropping time were obtained by means of shadowgrams of the light output from two of the light channels of the translator. Each shadowgram comprised the operation of the 10 test cards in sequence. Samples of these shadowgrams are shown on Figs. 1 and 2. For the purpose of these experiments the card dropping time is defined as the time from the release of the pull-up magnets until the full closure of the light channels of the translator exclusive of any card rebound. The data from the various experiments were tabulated and are given in Tables V to XI, inclusive.

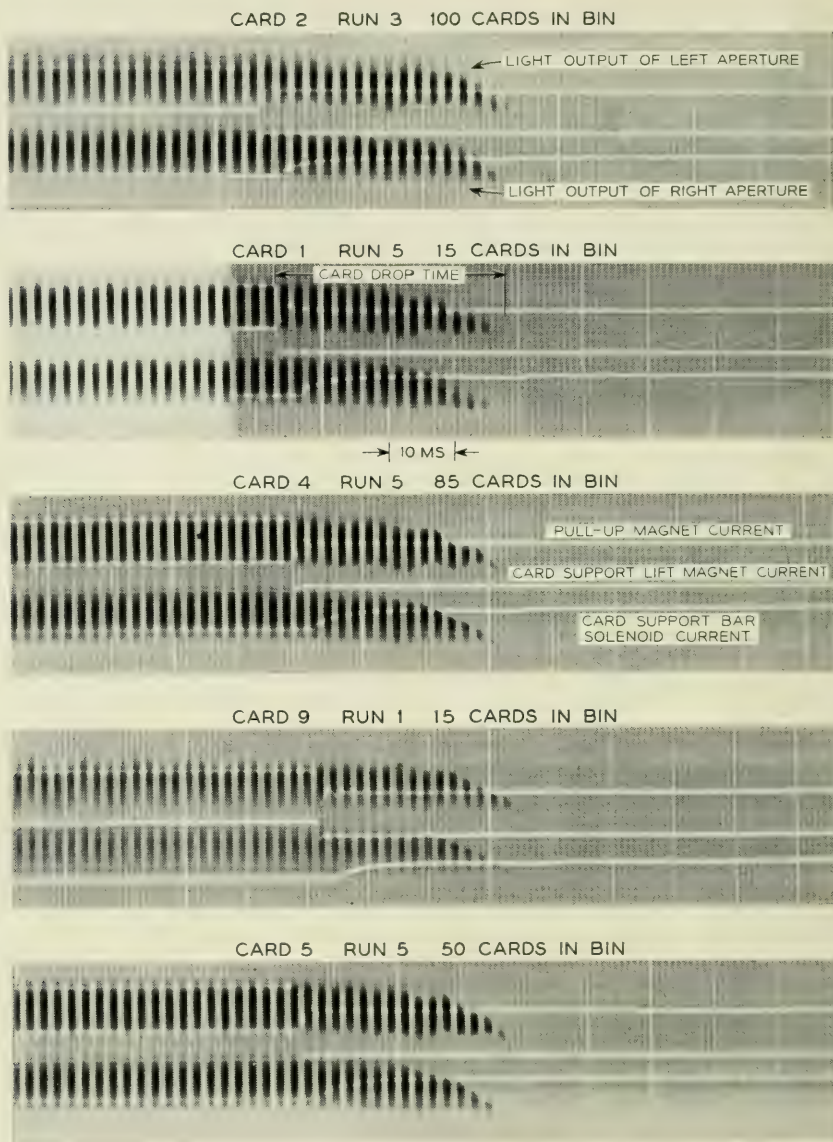
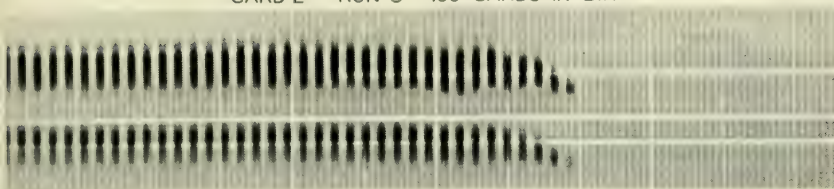
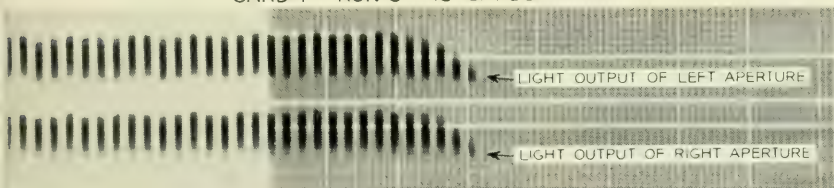


FIG. 1 — Card motion shadowgraphs, card support bars working.

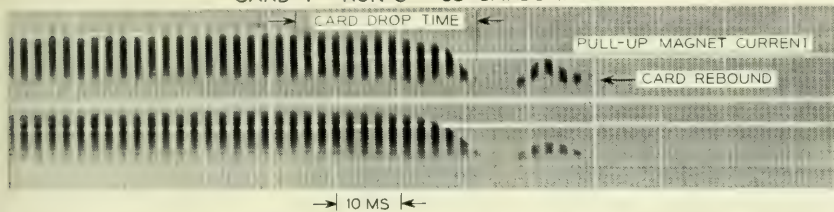
CARD 2 RUN 3 100 CARDS IN BIN



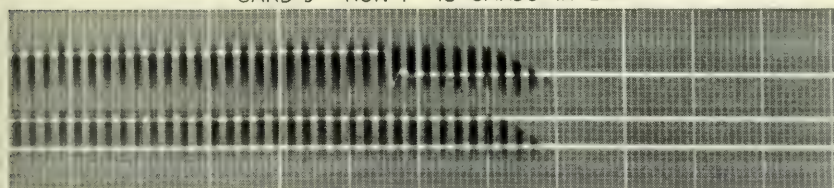
CARD 1 RUN 5 15 CARDS IN BIN



CARD 4 RUN 5 85 CARDS IN BIN



CARD 9 RUN 1 15 CARDS IN BIN



CARD 5 RUN 5 50 CARDS IN BIN

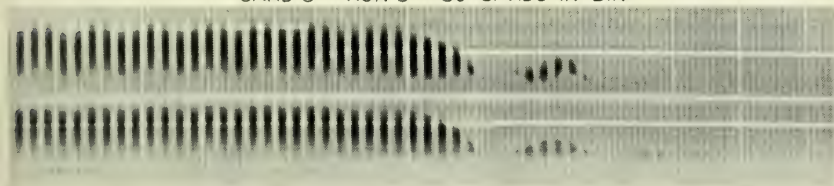


FIG. 2 — Card motion shadowgraphs, card support bars out.

TABLE IV — CARD TRANSLATOR BALANCED VERSUS UNBALANCED LOAD TEST — ARRANGEMENT OF CODED CARDS IN BIN I

No. of Cards	Arrangement of Coded Cards in Bin I																			
	← Left end of Bin															Right end of Bin →				
100	*1	*2	10	*3	15	*4	15	*5	10	*6	15	*7	15	*8	10	*9	*10			
50	*1	*2	5	*3	6	*4	7	*5	4	*6	7	*7	6	*8	5	*9	*10			
33	*1	*2	3	*3	4	*4	3	*5	3	*6	3	*7	4	*8	3	*9	*10			
16	*1	*2	1	*3	1	*4	1	*5	0	*6	1	*7	1	*8	1	*9	*10			

Note: Numbers preceded by * represent the test cards. Numbers in italics represent quantity of non-test coded cards placed between two consecutive test cards.

IV. ANALYSIS OF THE DATA

After checking the raw data for recording errors, they were analyzed and reduced in three distinct steps as follows:

1. Using the techniques of the analysis of variance* each variable and each measured interaction of several variables was tested as a possible assignable cause of variation.

2. Using the components of variance analysis† on those variables and combinations of variables found to be assignable causes in Step 1 their contributions to the overall variation was estimated, and

3. After tabulating the arithmetic mean for each value of the variables an upper and lower bounds were determined which estimates the allowance, or probable range of means, for similar experiments and operations on card translators having the same residual σ . The technique used was proposed and formulated by J. W. Tukey.³

The analysis of variance tables were reduced to the summary in Tables XII and XIII for the seven variable studied.

Proceeding to the tabulation of the means and the calculation of the allowances appropriate to these means, the data was reduced as in Table XIV. The magnitude of the several effects can now be noted. The effect of Idler load is quite meaningful — as the load on the machine is increased to the normal load of 85 we see a decrease in mean dropping time from 36.5 to 34.4 milliseconds. Also the slight increase for the overload of 100 is not significant statistically or engineeringwise. At first glance the results of Idler loads of 0, 50, 85 and 100 might seem to be inconsistent with those of Operating Loads of 15, 50, 85, 100, and 105. The mean dropping time of the Load of 15 (say) is the average dropping time of

* See Appendix for Discussion and References 2, 4 and 5.

† See Appendix for Discussion and Reference 6.

TABLE V CARD DROP TIME IN MILLISECONDS — NO CARDS IN X BINS CARD SUPPORT BARS OPERATING

		Bin # I		Bin # II		Bin # III		Bin # IV		Bin # V	
Card # →		1	6	2	7	3	8	4	9	5	10
Run #1	A	36	38	36	35.5	37	36.5	37	37.5	37.5	40
	B	36.5	38	36.25	37.5	35	37	36.5	37	37	39.5
	C	35.25	37.5	35	36	37	36	36	36	37.5	39
	D	35.5	38	37	36.25	36.5	36.5	36.5	36.5	37	39
	X	35.8	37.8	36.06	36.32	36.37	36.5	36.5	36.75	37.25	39.37
	R	1.25	0.5	2.0	2.0	2.0	1.0	1.0	1.5	0.5	1.0
Run #2	A	34	35	34	33.5	34	34	36	35	37	39
	B	33.5	34.5	34	34	34	34	35.5	33.5	37	38
	C	34	34.5	33.5	34	35	34	36	35	36.5	38.5
	D	35.5	33	33.5	34.5	34.5	34	35.5	34.5	36.5	38
Run #2	X	34.25	34.25	33.75	34.0	34.37	34	35.75	34.5	36.75	38.37
	R	2.0	2.0	0.5	1.0	1.0	0	0.5	1.5	0.5	1.0
Run #3	A	36.5	37.5	38	36.5	37	37	37	38	37.5	39.5
	B	36	37	36.5	35	36	37	36.5	36.5	38	39
	C	36	36.5	37.5	35.5	36	35	37	36.5	38	39.5
	D	35.5	36.5	37.5	35.5	36	35.5	35.5	36.5	37.5	38.5
Run #3	X	36.0	36.37	37.37	35.62	36.25	36.12	36.5	36.37	37.75	39.12
	R	1.0	1.0	1.5	1.5	1.0	2.0	1.5	1.5	0.5	1.0
Run #4	A	37	37	35.5	37	36	37	37	37.5	38	38.5
	B	36	37.5	35.5	36	36	37	36.5	37.5	38	38.5
	C	36	37.5	35	36	35.5	36	37	38	37.5	38
	D	35.5	37	36	36	36	36	37	37.5	37.5	37.5
Run #4	X	36.12	37.25	35.50	36.25	35.87	36.50	36.87	37.62	37.75	38.12
	R	1.5	0.5	1.0	1.0	0.5	1.0	0.5	0.5	0.5	1.0
Run #5	A	36	36	37	37	35.5	36.5	37	36.5	38.5	38.5
	B	36	35.5	37	37	37	37	37	37	38.5	39
	C	36	36	36	36	36	36.5	37	37	38.5	39
	D	36	35	35.5	36	35	36	36.5	37	38	39
	X	36	35.62	36.37	36.50	35.87	36.50	36.87	36.87	38.37	38.87
	R	0	0.5	1.5	1.0	2.0	1.0	0.5	0.5	0.5	0.5

Cards in Bins loaded with 15 cards in the presence of 4 bins loaded with 50, 85, 100, and 105 cards respectively and of bins loaded as a group with 0, 50, 85, or 100 cards. On the other hand the mean dropping time attributable to an Idler load of 50 cards (say) is the average of all dropping cards in the operating bins when the 7 Idler loads are 50 cards. Hence the statistical conclusion that the effect of loads in the operating bins is slight over the range of loads of 15 to 105 cards is reached, and that the

TABLE VI — CARD DROP TIME IN MILLISECONDS — 50 CARDS IN X BINS — CARD SUPPORT BARS OPERATING

		Bin #I		Bin #II		Bin #III		Bin #IV		Bin #V	
Card # →		1	6	2	7	3	8	4	9	5	10
Run #1	A	35	36	34	35	34.5	35	33	34	33.5	35
	B	33.5	35	34.25	34	33	33.5	32.5	33.5	33.5	34.5
	C	34.5	34	32.5	34	32.5	32.5	32	33.25	32	34
	D	34	32.5	32	34	33.5	33.5	33.25	32.5	33	33
	$\bar{X}$	34.25	34.37	33.19	34.25	33.37	33.62	32.69	33.31	33	34.12
	R	1.5	3.5	2.25	1.0	2.0	2.5	0.75	1.5	1.5	2.0
Run #2	A	35.5	35	34	34	32.5	35	35	35.5	35	35.5
	B	35.5	35.5	34	33.5	34.5	34.5	35.5	35	35.5	35
	C	35	35	34	35	34.5	34	36	35	34.5	34
	D	34	35	34	34	32.5	34	35.5	34	35	35
	$\bar{X}$	35.0	35.1	34	34.1	33.5	34.4	35.5	34.9	35.0	34.9
	R	1.5	0.5	0	1.5	2.0	1.0	1.0	1.0	1.0	1.5
Run #3	A	35	35	37.5	34	34.5	34.5	34	34	34	34.5
	B	36	36.5	37	36	35.5	35.5	34	34.5	34.5	36
	C	35	35.5	39	35	35.5	35	34.5	34.5	34.5	35.5
	D	35.5	37	39	35	35.5	34	34.5	34.5	34.5	35.5
	$\bar{X}$	35.37	36.0	38.12	35.0	35.25	34.75	34.25	33.37	33.37	35.37
	R	1.0	2.0	2.0	2.0	1.0	1.5	0.5	0.5	0.5	1.5
Run #4	A	37	37	35.5	35.5	36.5	35	35	35.5	36	37
	B	36.5	37	36	36.5	36.5	36	35.5	35.5	36	37
	C	37.5	38	35.5	36	36.5	35	36	35.5	35.5	36.5
	D	37	37.5	35.5	37.5	36.5	36.5	36.5	36	37	38
	$\bar{X}$	37	37.37	35.62	36.37	36.50	35.62	35.75	35.62	36.12	37.12
	R	1.0	1.0	0.5	2.0	0	1.5	1.5	0.5	1.5	1.5
Run #5	A	36	36	35.5	36	35.5	36	34.5	36	34.5	35
	B	35	35.5	35	36	35	35	34	34	34.5	35
	C	35.5	36	35.5	35.5	34.5	35	34	35	34.5	35
	D	36	35.5	35	35.5	34.5	35.5	34	34.5	34	34.5
	$\bar{X}$	35.62	35.75	35.25	35.75	34.87	35.37	34.12	34.87	34.37	34.89
	R	1.0	0.5	0.5	0.5	1.0	1.0	0.5	1.5	0.5	0.5

improvement in dropping time caused by increasing the overall load is generally consistent over the test bin loads.

One last estimate must be made — that of the σ' for a single measurement where*

$$\sigma'^2 = \sigma_s'^2 + \sigma_e'^2 + \sum \sigma_{\text{assignable causes}}'^2$$

* See Appendix for meaning of symbols and discussion.

TABLE VII — CARD DROP TIME IN MILLISECONDS — 85 CARDS IN X
BINS — CARD SUPPORT BARS OPERATING

		Bin #I		Bin #II		Bin #III		Bin #IV		Bin #V	
Card # →		1	6	2	7	3	8	4	9	5	10
Run #1	A	34.5	34	32.5	32.5	33	34	32.5	34	33.5	33.5
	B	35	34	34	32	32	34	33	33	33.5	33
	C	33	34	33.5	33.5	32.5	33.5	33	33.5	33.5	33
	D	35	34	34	32	33	33.5	33	33.5	32.5	34
	$\bar{X}$	34.4	34	33.5	32.5	32.9	33.7	32.9	33.5	33.25	33.4
	R	2.0	0	1.5	1.0	1.5	0.5	0.5	1.0	1.0	1.0
Run #2	A	33	34.5	32	34	33	34.5	34	33.25	33.5	33.5
	B	32	35	34.5	34.5	32	33.5	35	34	33.75	34
	C	34	35	33.5	34.5	33	34.5	35.5	35	35	35
	D	34	34	34	34.5	33.5	35	35	34.5	35	33
	$\bar{X}$	33.25	34.62	33.5	33.4	32.9	34.4	34.19	33.55	34.31	33.79
	R	2.0	1.0	2.5	0.5	1.5	1.5	1.5	1.75	1.5	1.5
Run #3	A	35	35	36.5	32.5	34	34.5	33	34.5	34	34
	B	35	35.5	37	34.5	35.5	34	33	34.5	34.5	34
	C	35	34.5	36	34.5	35.5	33	33.5	34	34.75	35
	D	35	35	37	34.5	35	34.5	34	34.5	34	34.5
	$\bar{X}$	35	35	36.62	34	35	34	33.37	34.37	34.31	34.37
	R	0	1.0	1.0	2.0	1.5	1.5	1.0	0.5	0.75	1.0
Run #4	A	37	37	36.5	36.5	36	36	35	35	36.5	36.5
	B	38	37.5	37	36	36.5	37	36	35	36.5	36.5
	C	37	37	37	35	36	36.5	35.5	35	36	36.5
	D	36	35	36	36	36	36	35	35.5	35.5	36.5
	$\bar{X}$	37	36.62	36.62	35.87	36.12	38.37	35.37	35.12	36.12	36.25
	R	2.0	2.5	1.0	1.5	0.5	1.0	1.0	0.5	1.0	1.0
Run #5	A	35	35	34	35	33.5	34	34	33.5	33.5	34
	B	34.5	34.5	34	34	34	33.5	34	33	33.5	33.5
	C	34	34.5	34	34.5	34	33.5	33	32.5	33	34
	D	34	35	33.5	34.5	33.5	34	33	33.5	32.5	34
	$\bar{X}$	34.37	34.75	33.87	34.50	33.75	33.75	33.50	33.12	33.12	33.87
	R	1.0	0.5	0.5	1.0	0.5	0.5	1.0	1.0	1.0	0.5

If the assignable causes contributing to this estimate are not removed then this σ' is the well known parameter for control charts.

The variable Runs may be an experimental variable which will not occur in operational usage since coded cards only will be used in the machines. The joint effect, however, of this variable with all the others is slight, for

$$\sigma'_{w0} = 2.68 \text{ ms without Runs, and}$$

$$\sigma'_{w} = 2.98 \text{ ms with Runs included.}$$

TABLE VIII—CARD DROP TIME IN MILLISECONDS—100 CARDS IN X BINS—CARD SUPPORT BARS OPERATING

		Bin #I		Bin #II		Bin #III		Bin #IV		Bin #V	
Card #→		1	6	2	7	3	8	4	9	5	10
Run #1	A	34	35.5	34.5	35	34	35	33.5	34	34.5	35
	B	34	35	35	35	34.5	34.5	34.5	35	35	34
	C	34	35	34	33	35	34.5	33.5	33.5	35	34
	D	35	33.5	34.5	35	35	34	35	33.5	33.5	34.5
	X	34.75	34.75	34.5	34.5	34.62	34.5	34.12	34.0	34.5	34.37
	R	1.0	2.0	1.0	2.0	1.0	1.0	1.5	1.5	1.5	1.0
Run #2	A	34	36.5	34.5	35.5	33.5	34.5	35.5	36	35	35.5
	B	34.5	35	35	35.5	33.5	35	35.5	36.5	35.5	34
	C	34	34	34	33	34	33.5	34	34	33.5	33
	D	33	32	33.5	32.5	33.5	33.5	34.5	34	33	34
	X	33.9	34.37	34.25	34.12	33.6	34.12	34.9	35.12	34.25	34.12
	R	1.5	4.5	1.5	3.0	0.5	1.5	1.5	2.5	2.5	2.5
Run #3	A	34.5	34.5	36.5	33.5	35.5	33.5	34	34	34	34
	B	35	34.5	38	34	34.5	33.5	34	33.5	33	34
	C	35	35.5	38.5	34	34.5	34.5	33.5	33.5	34.5	34.5
	D	34.5	35	39.5	34	35	35	34	34	34	35
	X	34.75	34.87	38.12	33.87	34.87	34.12	38.87	33.75	33.87	34.37
	R	0.5	1.0	3.0	0.5	1.0	1.5	0.5	0.5	1.5	1.0
Run #4	A	37	37	35.5	35.5	36	35.5	35	35.5	36	37
	B	36	36	37	36	36	35.5	35	35.5	37	36
	C	36	36.5	35.5	37	36	36.5	35	36.5	36	37
	D	37	36	35.5	35.5	35.5	35.5	34.5	35	36	37
	X	36.5	36.37	35.87	36	35.87	35.75	34.87	35.62	36.25	36.75
	R	1.0	1.0	1.5	1.5	0.5	1.0	0.5	1.5	1.0	1.0
Run #5	A	33.5	34	33.5	33.5	33.5	33.5	32.5	33.5	33.5	33
	B	33.5	33	34	34.5	34	34	32.5	33.5	33.5	33.5
	C	35	35	34	34.5	33.5	34	33	33.5	33.5	33
	D	34	34	34.5	33.5	34	34	33	33	33	34
	X	34	34	34	34	33.75	33.87	32.75	33.37	33.37	33.37
	R	1.5	2.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	1.0

The overall estimate of Mean dropping time was found to be 35.1 ms. We can predict therefore that the dropping time of a card chosen at random from a normal translator at the beginning of its life will be $35.1 \pm 3\sigma'$ (without Runs) or between 27 and 43 ms from the well known 3σ limits.

When the analysis of variance with the card support bars out is examined it is found that the sampling error σ_s^2 remains stationary but

TABLE IX — CARD DROP TIME IN MILLISECONDS — 85 CARDS IN X
BINS — CARD SUPPORT BARS OUT

		Bin * I		Bin * II		Bin * III		Bin * IV		Bin * V	
Card # →		1	6	2	7	3	8	4	9	5	10
Run #1	A	23.5	27	26.5	26	28	28	25	24	26	26.5
	B	23.5	28	25.5	26.5	28	27	25	24	25.5	26
	C	23	27	26	26	29	28	25	23	26	26.5
	D	24	26.5	26.5	27	29.5	27.5	25	24	26.25	26
	X	23.5	27.1	26.1	26.4	28.6	27.6	25	23.75	25.94	26.25
	R	1.0	1.5	1.0	1.0	1.5	1.0	0	1.0	1.75	0.5
Run #2	A	24.5	26.5	26	23.5	22	21.5	34	31.5	32	27.5
	B	25	27	25	22	21.5	22	33.5	31	32	28
	C	24	27.5	26.5	23.5	21.5	23	34.5	31.5	31.5	27.5
	D	24.5	26.5	25	22.5	22	22	34	31.5	33	28
	X	24.5	26.9	25.6	22.9	21.75	22.12	34.0	31.4	34.6	27.75
	R	1.0	1.0	1.5	1.5	0.5	1.5	1.0	0.5	1.5	0.5
Run #3	A	28	29	37.5	28	34	28	29	28.5	24	22
	B	27	29	37	28	33	28	30	28	24	21
	C	28.5	30	37.5	29	32.5	28	28.5	28	23.5	20
	D	28	30	38	29	33	28	29	27.5	24	21
	X	27.87	29.5	37.5	28.5	33.12	28	34.12	28	23.87	21
	R	1.5	1.0	1.0	1.0	1.5	0	1.5	1.0	0.5	2.0
Run #4	A	16.5	23.5	22	24	28	29	28.5	32	34	34
	B	17	23	22	23.5	28	29	28	33	33.5	33.5
	C	17	23	21	22	28	29	27.5	32	30.5	33
	D	17	23.5	22	22.5	28	29	28	31	33.5	36.5
	X	16.87	23.25	21.75	23	28	29	28	32	32.87	33.50
	R	0.5	0.5	1.0	2.0	0	0	1.0	2.0	3.5	3.0
Run #5	A	16.5	17	25	25.5	28.5	28	29	30	26	25.5
	B	16	17	26	26.5	28	28	29	29.5	26	25
	C	18	17.5	25	25.5	28	28	28.5	29.5	26	25
	D	16.5	17.5	25	25.5	28	28	29	30	26	25.5
	X	16.75	17.25	25.25	25.75	28.12	28	28.87	29.75	26	25.25
	R	2.0	0.5	1.0	1.0	0.5	0	0.5	0.5	0	0.5

the experimental error is inflated tenfold. The assignable causes found in the previous experiment are found here, where measured, with one exception — the interaction of Codes on Positions. The estimate of σ' for a single reading with the card support bars out now becomes

$$\sigma' = 8.56 \text{ ms.}$$

Since the overall estimate of mean dropping time with the card support

TABLE X — TILT STUDY DATA — CARD DROP TIME IN MILLISECONDS —
CARD SUPPORT BARS OUT

		Bin #I		Bin #II		Bin #III		Bin #IV		Bin #V	
Card #→		6	1	2	7	8	3	4	9	10	5
Tilt #0	A	21.5	19.5	29	26	30.5	27.5	32.5	32.5	31	29
	B	21.5	19.5	31	25.5	32	28.5	31.5	35	30.5	28.5
	C	21.5	19.0	29.5	26	31	28	31	34	30.5	29
	D	21.5	19.0	29.5	25.5	30.5	28.5	32	34.5	30.5	29.5
	Σ	86	77	119	103	124	112.5	127	136	122.5	116
	$\bar{X}$	21.5	19.25	29.75	29.75	31	28.13	31.75	34	30.63	29
	R	0	0.5	2	0.5	1.5	1	1.5	2.5	0.5	1.0
Tilt #1	A	27.5	21	38	28	38.5	28.5	44	28.5	47	31.5
	B	25	21	37	27.5	38	29	44	29	49	31
	C	25.5	20.5	38	27.5	43	28.5	47	29.5	46	31
	D	25.5	21.5	39	28	42.5	28	43.5	29	46.5	32
	Σ	103.5	84	152	111	162	114	178.5	116	188.5	125.5
	$\bar{X}$	25.88	21	38	27.75	40.5	28.5	44.63	29	47.13	31.38
	R	2.5	1	2	0.5	5	1	3.5	1	3	1
Tilt #2	A	29	23.5	62.5	32	67	32	73	32.5	92.5	38.5
	B	38.5	23	57.5	31	63.5	31.5	72.5	32	62.5	37
	C	33	22.5	58.5	30.5	59	31	68.5	32	66	37.5
	D	33	23	59	31.5	59.5	31.5	75	33	67.5	37.5
	Σ	132.5	92	237.5	125	249	126	289	129.5	288.5	150.5
	$\bar{X}$	33.13	23	59.38	31.25	62.25	31.5	72.25	32.38	72.13	37.63
	R	9.5	1	5	1.5	8	1	6.5	1	30	1.5
Tilt #3	A	58	23	68	33.5	109.5	34	Did not operate	34.5	69.5	39.5
	B	41.5	23.5	69	33	95	33.5		34.5	91.5	38.5
	C	34	23	80	33.5	77	34		36	87	39.5
	D	31.5	24.5	80.5	34	86.5	35		36	95	39
	Σ	165	94	297.5	134	368	136.5		141	343	156.5
	$\bar{X}$	41.25	23.5	74.38	33.5	92	34.13		35.25	85.75	39.13
	R	26.5	1.5	12.5	1	32.5	1.5		1.5	25.5	1

bars out was 26.7 ms, it can be predicted that the limits for the dropping time for a single card at the onset of life of a card translator will be 1 and 52 ms. It is to be noted that while the upper limits in both these estimates are reasonably close, the lower limits are quite different. This is to be expected since the card support bars may delay the dropping of a card and thus restrict operations that might be fast.

Tilt Study Analysis

A fundamental assumption underlying any comparison is that the compared elements have been measured with the same precision. Scrut-

TABLE XI — BALANCED VERSUS UNBALANCED LOAD — CARD DROP
TIME IN MILLISECONDS

Treatment	Load	Look	Test Card									
			1	2	3	4	5	6	7	8	9	10
W	600	A	24.5	21.5	25	21.5	28.5	28	26.5	29	30.5	31.5
		B	24	22	26	27.5	28	27.5	27	29	29.5	31
	400	A	24	22	27.5	29.5	28.5	29	28	30	29	34
		B	24	22.5	28	28	28	28	28	30	30.5	33
	200	A	26	24.5	29.5	30	30	29	29	30	31	33
		B	25.5	24	30	30	30	29.5	28.5	29.5	30	33.5
X	600	A	25	22.5	27	28	27.5	28	27.5	26	30	32
		B	24.5	23	27	28.5	28	28	28	29	30	33
	400	A	25	24.5	29	29	29.5	28	28.5	30	30	33
		B	25	24.5	29.5	29	29.5	28	28	29	30.5	33
	200	A	28	26	29.5	29	29	28.5	27.5	27.5	24.5	26
		B	27	25.5	29.5	29	28.5	27.5	26.5	28	25	26
Y	600	A	24.5	22.5	27	28	28.5	28.5	28	29	30	34
		B	24	23	27	28	28	28	28.5	30	29	32.5
	400	A	25.5	23	29	29.5	29	28	28.5	28.5	28.5	33.5
		B	25	24	29.5	29.5	29	28	28.5	29	29.5	33
	200	A	27.5	26.5	28.5	28.5	27.5	27.5	26	27	25	26
		B	27.5	26	28.5	28.5	27.5	28	26	28	23.5	26
Z	600	A	21	20	23	23.5	25	25.5	26	25.5	26	26.5
		B	21.5	20	23	24.5	25.5	26.5	25.5	25.5	26.5	26.5
	400	A	18	18.5	23	26.5	27	25	24.5	24.5	24.5	27
		B	19	18.5	22	26	27	25.5	25.5	25	25	26
	200	A	19.5	18.5	20	21.5	22	22	22	22	22	22
		B	18	18.5	21	22	22.5	22	21.5	21.5	22.5	22

tiny of the data cast grave doubt on this assumption and each card position was analyzed separately as shown below.

CARD POSITION A ANALYSIS OF VARIANCE

Source	Degrees of Freedom	Sum of Squares	Mean Square
Bins	4	8947	2238
Tilt	3	25653	8551
(Expt. Error) Bins x Tilt*....	11	2882	262
Total	18	37482	

* One observation was lost, and Fisher's Missing Plot Technique was used (5).

CARD POSITION B ANALYSIS OF VARIANCE

Source	Degrees of Freedom	Sum of Squares	Mean Square
Bins	4	1515	379
Tilt	3	492	164
(Expt. Error) Bins x Tilt.....	12	168	14
Total	19	2175	

The two estimates of experimental error being 262 and $14 \overline{ms}^2$ respectively against a previous estimate (Table XIII) on 24 degrees of freedom of $10.9 \overline{ms}^2$, we must accept Position A as coming from a different universe with respect to position B and the previous seven variable experiments. Further Position B on the same evidence has been measured with equivalent precision when compared to the previous study with the card support bars out.

The individual dropping times of Position B cards over the experiment range from 19 to 39.5 ms., well within the predicted dropping times. The mean dropping times for each tilt of card Position B are as follows:

Tilt	0	1	2	3
	27.2 ms	27.5 ms	31.2 ms	33.8 ms

By contrast the mean dropping times for Position A are:

Tilt	0	1	2	3
	28.9 ms	39.2 ms	59.9 ms	73.6 ms

with a range of 21 ms to 109.5 ms on the fourth tilt.

TABLE XII — ANALYSIS OF VARIANCE — SUMMARY —
WITH CARD SUPPORT BARS

Source	Degrees of Freedom	Sum of Squares	Mean Squares	Mean Squares is an Estimate of*
1. Looks	3	5.65	1.88 N.S.(+)	
2. Code	1	4.46	4.46†	$\sigma_e^2 + 20(\sigma_{eb}^2 + \sigma_{cp}^2 + \sigma_{cl}^2 + \sigma_{cr}^2) + 100\sigma_e^2$
3. Idlers	3	520.48	173.49‡	$\sigma_e^2 + 10(\sigma_{ib}^2 + \sigma_{ir}^2) + 25\sigma_{ei}^2 + 50\sigma_i^2$
4. Bins	4	53.43	13.35‡	$\sigma_e^2 + 10\sigma_{ib}^2 + 20\sigma_{cb}^2 + 40\sigma_b^2$
5. Runs	4	359.64	89.91‡	$\sigma_e^2 + 20\sigma_{cr}^2 + 10\sigma_{ir}^2 + 40\sigma_r^2$
6. Loads	4	26.07	6.52‡	$\sigma_e^2 + 20\sigma_{cl}^2 + 40\sigma_l^2$
7. Positions	4	35.96	8.99‡	$\sigma_e^2 + 20\sigma_{cp}^2 + 40\sigma_p^2$
8. Codes x Idlers	3	6.58	2.19 N.S.	$\sigma_e^2 + 25\sigma_{ei}^2$
9. Codes x Bins	4	20.34	5.08‡	$\sigma_e^2 + 20\sigma_{cb}^2$
10. Codes x Runs	4	16.78	4.20‡	$\sigma_e^2 + 20\sigma_{cr}^2$
11. Codes x Loads	4	11.99	3.00†	$\sigma_e^2 + 20\sigma_{cl}^2$
12. Codes x Positions	4	25.47	6.37‡	$\sigma_e^2 + 20\sigma_{cp}^2$
13. Idlers x Bins	12	162.52	13.54‡	$\sigma_e^2 + 10\sigma_{ib}^2$
14. Idlers x Runs	12	203.99	17.00‡	$\sigma_e^2 + 10\sigma_{ir}^2$
15. Experimental Error	136	128.76	0.95	σ_e^2
16. Sampling Error	597	224.88	0.39	σ_s^2
	799	1807.00		

* See Appendix and Reference.

† Significant at 5 per cent level.

‡ Significant at 1 per cent level.

(+) N.S. = Not significant at 5 per cent level.

Clearly, Card Position A gives rise to an undesirable assignable cause which must be dealt with, and Card Position B while showing tilt as an assignable cause has a dropping time at the extreme tilt well within the allowable tolerances (see Conclusions).

Balanced Loading

The two main variables were total machine load in numbers of cards (Number) and distribution of cards over bins (Loading). The results of this experiment were so clear that little analysis was necessary. Although the mean dropping time of cards distributed uniformly over the twelve bins was statistically significantly different from the means of cards from the three nonuniform distributions the magnitude of the difference 4.8 ms was not sufficiently large to cause concern. Furthermore a decrease in mean dropping time was observed as the total load in-

TABLE XIII — ANALYSIS OF VARIANCE — SUMMARY —
WITHOUT CARD SUPPORT BARS

Source	Degrees of Freedom	Sum of Squares	Mean Square	Mean Square is an Estimate of*
1. Looks	3	1.72	0.57 N.S.	
2. Code	1	3.43	3.43 N.S.	$\sigma_e^2 + 5\sigma_{cb}^2 + 5\sigma_{cr}^2 + 5\sigma_{cl}^2 + 5\sigma_{cp}^2 + 25\sigma_c^2$
3. Idlers				
4. Bins	4	714.69	178.67†	$\sigma_e^2 + 5\sigma_{cb}^2 + 10\sigma_b^2$
5. Runs	4	275.61	68.90†	$\sigma_e^2 + 5\sigma_{cr}^2 + 10\sigma_r^2$
6. Loads	4	1695.52	423.88†	$\sigma_e^2 + 5\sigma_{cl}^2 + 10\sigma_l^2$
7. Position	4	239.42	59.86†	$\sigma_e^2 + 5\sigma_{cp}^2 + 10\sigma_p^2$
8. Codes x Idlers				
9. Codes x Bins	4	205.39	51.35†	$\sigma_e^2 + 5\sigma_{cb}^2$
10. Codes x Runs				
11. Codes x Loads				
12. Codes x Positions	4	144.09	36.02†	$\sigma_e^2 + 5\sigma_{cp}^2$
13. Idlers x Bins				
14. Idlers x Runs				
15. Experimental Error	24	261.28	10.89	σ_e^2
16. Sampling Error	147	48.86	0.33	σ_s^2
	199	3689.01		

* See Appendix and Reference 6.

† Significant at 5% level.

‡ Significant at 1% level.

(+) N.S. = Not significant at 5% level.

creased regardless of the distribution of cards, confirming a conclusion from the first experiment.

V. CONCLUSIONS OF OVER-ALL STUDY

When considered in the normal cycle of translator operation with the card support bars functioning, the range of card drop times was observed to be from 32 to 40 ms with a mean of 35.1 ms. Based on the results obtained in this study it is predicted the dropping time of a card chosen at random in a new translator will be between 27 and 43 ms. This includes all the known variables except the life of the cards. The test on this variable is continuing and all conclusions reached in this report may be modified somewhat by the results of this life study which will be reported at a later date. The relative effect of the several variables is tabulated in Table XIV. The only two variables found to have any sig-

TABLE XIV — MEAN CARD DROPPING TIMES WITH ALLOWANCES
SUMMARIZED WITH REGARD TO VARIABLES

Variable	No. Readings in each Mean	Mean Dropping Time Milliseconds				Range of Means	Allowance
Looks		A	B	C	D		
C.S. In...	200	35.2	35.2	35.1	34.9	Negligible	Negligible
C.S. Out...	50	26.8	26.6	26.6	26.8		
Codes		3 digit	6 digit				
C.S. In...	400	35.0	35.2			Negligible	Negligible
C.S. Out...	100	26.6	26.8				
Idlers		0 cards	50	85	100		
C.S. In...	200	36.5	35.0	34.4	34.6	2.1	±0.4
Graeco-Latin Square							
Runs		1	2	3	4	5	
C.S. In...	160	34.6	34.5	35.3	36.3	34.8	1.8 ±0.5
C.S. Out...	40	26.0	26.9	28.6	26.9	25.1	3.5 ±3.0
Bins		I	II	III	IV	V	
C.S. In...	160	35.3	35.1	34.8	34.8	35.5	0.7 ±0.5
C.S. Out...	40	23.4	26.3	27.4	29.0	27.5	5.6 ±3.0
Positions		ab	ad	ce	dg	fg	
C.S. In...	160	35.5	35.3	34.9	35.0	34.9	0.6 ±0.5
C.S. Out...	40	28.9	26.1	26.2	26.0	26.5	2.9 ±3.0
Loads		15 cards	50	85	100	105	
C.S. In...	160	34.9	35.0	35.0	35.4	35.3	0.5 ±0.5
C.S. Out...	40	21.6	25.3	29.1	29.3	28.3	7.7 ±3.0
Graeco-Latin Sq. & Codes		Min		Max.			
C.S. In...	16	33.6		37.6		4.0	±2.5
C.S. Out...	4	16.8		37.5		20.7	±8.4
Idlers & Bins							
C.S. In...	40	34.0		38.2		4.2	±1.6
Idlers & Runs							
C.S. In...	40	33.4		36.8		3.4	±1.6

nificance when operating in the normal manner were the load in the machine and the number of coded and uncoded cards in the working bin. In a lightly loaded machine the flux from the pull-up magnet is concentrated in the few cards and therefore in each card there is a greater amount of flux to decay before the card is free to drop. When a bin is loaded with all coded cards the dropping time may be increased because the cards are slightly deformed by the coding process which in turn increases the effective thickness of the cards thus reducing their freedom in the bin.

Although these two variables were significantly large they still are included in the range of dropping times mentioned above and are not of sufficient magnitude to warrant special consideration in loading the

TABLE XV — CARD TRANSLATOR — SUMMARY OF DATA ON CARD REBOUND — CARD SUPPORT BARS OUT

No. of operations.....	200	100%
No. of operations with no rebound.....	83	41.5%
No. of operations with one rebound.....	116	58%
No. of operations with two rebounds.....	1	0.5%
Maximum duration of one rebound.....	17 ms.	
Minimum duration of one rebound.....	6 ms.	
Duration of two rebounds.....	28 ms.	
Maximum magnitude of a rebound.....	$\frac{1}{3}$ reopening	

Bins	No. of Operations	No. of Operations with Rebound	Per cent of Operations with Rebound
1	40	28	70
2	40	24	60
3	40	25	62.5
4	40	15	37.5
5	40	25	62.5
Load	No. of Operations	No. of Operations with Rebound	Per cent of Operations with Rebound
15	40	15	37.5
50	40	28	70
85	40	23	57.5
100	40	23	57.5
105	40	28	70

translators. The effect of the load in the machine was further investigated in the balanced vs. unbalanced loading test and although it was found that a balanced load produced a faster drop time, the amount of improvement obtained by balancing the load is not enough to warrant special loading instructions.

The card dropping time as observed with the card support bars out of the circuit ranged from 16 to 40 ms with a mean of 26.7 ms and the predicted total range is from 1 to 52 ms. Although the minimum dropping time is considerably less with the card support bars out, the translator cannot be operated satisfactorily in this manner. The cards rebound on a majority of the operations; in fact, 58 per cent of the operations with the card support bars out had measurable rebound. When operated in the normal manner no card rebound was observed at any time. A summary of the amount of card rebound observed is given in Table XV.

As a result of this study, it was concluded that the requirement for minimum 65 cards per bin should be removed and that the field be permitted to place as few cards as they find convenient in any bin. A second conclusion was that the maximum number of cards per bin be set at 100. This will allow a total of 1,200 cards per machine, an increase of

20 per cent in the capacity of a translator over the design objectives. The requirement that an uncoded card be placed next to each separator should remain until the life test which is currently in progress is concluded. These uncoded cards are to be included in the 100 cards per bin figure. A preliminary analysis of the data indicates that it is probably necessary that this requirement be retained for field use.

With regard to the leveling of the machine, it was found that if the translator table is leveled consistent with the normal practice of the installation department no further leveling is necessary when the translator is installed on the table provided an uncoded card is next to each separator. When only cards that were at least one removed from the separator were considered it was found that the Translator could be tilted 1 inch in 3 feet without seriously affecting the card drop time.

Considering the ranges of the several variables considered and the results of the analysis of the data, it appears that there is no major unknown variable having an effect on the card dropping time. It is also believed that the results of the work on this machine can be considered representative of the results that will be obtained on another new production model translator.

APPENDICES

I. ANALYSIS OF VARIANCE

The general theory of the analysis of variance has been formulated and discussed at length by several authors.^{2, 4, 5} Basically it reduces to the concept that in any set of data obtained from a statistically designed experiment the total sum of squares of deviations from the mean can be partitioned into orthogonal components, and that under certain restrictions the distribution of each component falls into known patterns. Hence data taken from designed experiments can be examined for conformance to the known pattern, and a lack of conformity indicates an assignable cause of variation. Further, the distribution of the ratio of mean square deviations under specified conditions has been tabulated as the table of the F ratio. It has also been shown that when a treatment variable is not a parameter or assignable cause of variation in the experiment, the partitioned component for that variable must contain only residual variation. Thus, the analysis of variance tests the hypothesis that the treatment means for a given variable are all equal (i.e., the variable is not a parameter) by testing the ratio of the mean squares of mean deviations for the variable to the residual mean square, i.e., the F ratio. When this F ratio is larger than the critical value at the α^{th} level, the variable is said to be *significant at the α^{th} level*. That is, let

us assume a null hypothesis that the variable in question is not a parameter or assignable cause, and select a critical value, F^* , such that the probability of observing an F ratio greater than F^* (when the null hypothesis is true) is small (say 0.01). Then if the F ratio is computed from our experimental observations and the null hypothesis rejected when this ratio is larger than F^* , on the average incorrect decisions will be made not more than 1 per cent of the time. This method of evaluation is not trivial and in complex situations reference should be made to the literature or to experts.

II. COMPONENTS OF VARIANCE

A basic difference between the estimation of the Components of Variance⁶ and the Analysis of Variance above is the concept of the underlying model or law. The Analysis of Variance tests the hypothesis that the treatment variable is not a parameter. In the estimation of Components of Variance we assume that the observed effect of the several values of a given variable is a random sample from a normal population of effects from these values. If u_{iv} is the true effect of the i^{th} value of the v^{th} variable, then the component of variance due to the v^{th} variable, σ_v^2 , is found from

$$\sigma_u^2 = \frac{\sum_{i=1}^k \mu_{iv}^2 - \frac{\left(\sum_{i=1}^k \mu_{iv}\right)^2}{k}}{k - 1}.$$

If $u_{1v} = u_{2v} = \cdots = u_{kv}$, then $\sigma_v^2 = 0$, and the mean square for variable v contains only residual variation. If the variable v is a parameter, $\sigma_v^2 > 0$; and the mean square for variable v contains $\sigma_e^2 + M\sigma_v^2$ (M measurements being made at each of the k levels of the variable). It is desirable to estimate the component of variability of each variable, in order to be able to estimate the variability of a measurement which is affected by these variables. That is, if there are ρ variables whose components of variance are σ_i^2 , $i = 1, \cdots, \rho$ respectively and if the measurement, x , is influenced by all of these variables, then

$$\sigma_x^2 = \sum_{i=1}^{\rho} \sigma_i^2 + \sigma_e^2, \quad \text{and} \quad \sigma_x = \sqrt{\sum_{i=1}^{\rho} \sigma_i^2 + \sigma_{\text{error}}^2}.$$

Referring to Table XII and using the column of mean squares, we make the following inferences:

Since the ratio of mean square for variables to mean square for experimental error is "significantly" large for all the main variables, excluding Looks, these main variables are considered to be assignable causes of variation. It is also evident that the three digit cards are caus-

ing an effect quite different from that of the six digit cards when the four variables, Bins, Runs, Loads, and Position are allowed to vary. Since the ratio of the mean square of variables of lines 9, 10, 11, and 12 of Table XII to mean square for experimental error is significantly large, in the same way lines 13 and 14 show that the effect of Idler loads is not independent of the Bins and Runs effects. Statistically then, we have isolated many significant effects, but note that our experimental error, σ^2 , for samples of four readings is 0.95 ms^2 and σ is then 0.975 ms .

If we compare two means $\bar{x}_1$ and $\bar{x}_2$ each based on N samples of four observations each we will detect as significant, differences as small as

$$3\sigma_{\bar{x}_1 - \bar{x}_2} = \frac{3\sqrt{2\sigma}}{\sqrt{N}} \frac{4.095}{\sqrt{N}}.$$

When N is large we will, therefore detect as significant, differences which may be of no interest engineering-wise. Thus not only is the significance of the effects of interest but also the magnitude of the effect.

When the component of variance attributable to each of the significant effects is estimated only four are so large as to be of interest to the engineer. The four variables are Idlers, Runs, interaction of Idlers on Runs and Bins. The estimates of the components of variance, $\hat{\sigma}_{\text{effect}}^2$, are obtained by equating the linear combinations of the components of variance shown in the right hand column of Table XII to the mean squares which estimate them and solving.

The component estimates are:

	σ^2
Idlers	3. $\frac{\text{ms}^2}{\text{ms}^2}$
Runs	1.55 $\frac{\text{ms}^2}{\text{ms}^2}$
Idlers x Runs	1.60 $\frac{\text{ms}^2}{\text{ms}^2}$
Idlers x Runs	1.26 $\frac{\text{ms}^2}{\text{ms}^2}$

III. SOURCES AND MEASURES OF ERROR

In any experiment a decision must be made as to the number of experimental units to be measured and the number of repeated measurements to be made on each unit.⁷ It is important to note that measurements made on the same unit and a measurement made on each of several units give rise to two distinct sources of variation, and that both of these should be estimated. Consider making n measurements on each of k units, where the k units are a random sample of units belonging to a normal universe with mean μ and variance σ_b^2 . Further a set of measurements on the i^{th} unit is a random sample of measurements from a normal universe of measurements with mean μ_i and variance σ_w^2 . Clearly, if

the set contains only one measurement ($n = 1$), we cannot estimate σ_w^2 , and if there is only one unit ($k = 1$) we cannot estimate σ_b^2 . When both $n, k > 1$ we can estimate simultaneously both σ_b^2 and σ_w^2 . Let X_{ij} be the j^{th} measurement on the i^{th} unit,

$$j = 1, \dots, n; i = 1, \dots, k.$$

and $\bar{X}_i$ be the mean of the i^{th} unit,

$\bar{X}$ be the mean of all the units.

We can estimate σ_w^2 directly by computing

$$\sigma_w^2 = \frac{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{k(n-1)}, \quad \text{but}$$

σ_b^2 can only be estimated indirectly by first estimating $\sigma_w^2 + n\sigma_b^2$ from

$$\frac{n \sum (\bar{X}_i - \bar{X})^2}{k-1}$$

Then the estimate of σ_b^2 is

$$\frac{1}{n} \left(\frac{n \sum_{i=1}^k \frac{(\bar{X}_i - \bar{X})^2}{k-1}}{k-1} - \frac{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{(n-1)k} \right).$$

Since $\sigma_{\bar{x}}^2 = \frac{1}{k} (\sigma_b^2 + \frac{\sigma_w^2}{n})$, it is clear that if σ_b^2 is large relative to σ_w^2 , then

$\bar{X}$, for fixed $M = nk$, will have greater precision if k is large and n is small. The estimate of σ_w^2 is called the sampling variance or sampling attributable to repeated measurements. The estimate of σ_b^2 is called the component of variance due to experimental variation free of sampling error. For a given experiment the estimate of $\sigma_w^2 + n\sigma_b^2$ is called the experimental error term and measures the precision of measurement of a unit. In the experiment $\sigma_w^2 = \sigma_s'^2$, and $\sigma_b^2 = \sigma_e'^2$.

BIBLIOGRAPHY

1. L. N. Hampton and J. B. Newsom, "The Card Translator for Nationwide Dialing," B.S.T.J., **32**, pp. 1037-1098, Sept., 1953.
2. W. G. Cochran and G. M. Cox, "Experimental Designs", J. Wiley and Sons, New York, 1950.
3. J. W. Tukey, "Comparing Individual Means in the Analysis of Variance", Biometrics, Vol. 5, No. 2, 1949.
4. W. J. Dixon and F. J. Massey, "Introduction to Statistical Analysis", McGraw Hill, New York, 1951.
5. R. A. Fisher, "Design of Experiments", Oliver and Boyd, London, 1950.
6. S. Lee Crump, "The Estimation of Variance Components in the Analysis of Variance", Biometrics, Vol. 2, No. 1, 1946.
7. C. Eisenhart, "Assumptions Underlying the Analysis of Variance", Biometrics, Vol. 3, No. 1, 1947.

Wave Propagation Along a Magnetically-Focused Cylindrical Electron Beam

By W. W. RIGROD and J. A. LEWIS

(Manuscript received August 24, 1953)

This paper analyses the nature of wave propagation along a cylindrical electron beam, focused in Brillouin flow by means of a finite axial magnetic field. Two different types of conducting boundaries external to the beam are treated: (1) the concentric cylindrical tube, forming a drift region; and (2) the sheath helix, forming a model of the helix traveling-wave tube. The field solution of the helix problem is used to evaluate the normal-mode parameters of an equivalent circuit seen by a thin beam, thereby permitting computation of the gain constant of growing waves. The gain constant of the cylindrical beam with Brillouin flow is found to exceed that of a similar beam with rectilinear flow, presumably because of the transverse component of electron motion in the former.

INTRODUCTION

The theory of the helix traveling-wave has been treated in previous papers,¹⁻⁴ for cases in which the electrons move along straight lines parallel to the axis of the helix, as though immersed in an infinitely strong magnetic field. In practice, however, the electron beam is focused by a magnetic field of finite intensity,^{5, 6} such that the electrons follow spiral paths about the common axis. The purpose of this paper is to extend traveling-wave tube theory to the case of such focused beams, and to compare the gain constants for the two types of electron motion. The motion of the beam in an infinite field is usually described as rectilinear flow; that in a finite focusing field, as Brillouin flow.

The gain constant of the dominant mode in a traveling-wave tube may be computed from the field solution for the electron beam in the presence of its circuit structure. This procedure, however, requires the solution of cumbersome transcendental equations for each particular set of dimensions and operating conditions. A more flexible method of analysis has been provided by Pierce,¹ based on an expansion in terms of

normal modes of propagation. For any particular *type* of beam and circuit, three circuit parameters must be evaluated from the field solution. The performance of the traveling-wave tube is then described quite accurately by a cubic equation containing these parameters, over a wide range of dimensions and operating conditions. The usefulness of this normal-mode method has been further enhanced by publication of a nomograph⁷ for the calculation of the gain constant.

In its initial form, the normal-modes solution for a helix traveling-wave tube was greatly simplified by the assumption that the electron beam is so thin that the electric field acting on it is constant. Employing the field solution for a beam of finite thickness in a helix, Fletcher⁴ was able to compute the circuit parameters for the solid and hollow cylindrical electron beams, respectively, confined to rectilinear flow.

This procedure will now be extended to cylindrical beams in Brillouin flow, in which transverse electron motion occurs. First, it will be necessary to solve the field equations for this type of beam in a helix. As a by-product of this computation, the solution of the field equations for the beam in a concentric drift tube will briefly be given. Finally, with some restrictions, the helix parameters will be evaluated, and the gain of helix amplifiers with such beams compared with that obtained with otherwise identical rectilinear beams.

FIELD EQUATIONS IN THE ELECTRON BEAM

When a small ac field is impressed upon a short length of electron beam, the electrons respond by executing small ac excursions about their steady-state trajectories. These ac motions of charged particles constitute a transverse distribution of ac currents, which in turn excites an ac field distribution. The propagation of an ac signal along a beam depends upon the reciprocal action of these currents and fields.

To find the propagation constants for a particular configuration of electron stream and enclosure, we must therefore solve Maxwell's equations in the presence of the ac driving currents in the beam, subject to the external boundary conditions. When the fields and currents possess circular symmetry, these equations may be formally separated into TE and TM groups.² In addition, as we are concerned only with "slow" waves, the equations may be simplified by neglecting all terms of relative magnitude k^2/γ^2 , where k is the wave number in free space, and γ the propagation wave number.

TM WAVE

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial E_z}{\partial r} \right) - \gamma^2 E_z = \frac{\gamma^2}{j\omega\epsilon} J_z + \frac{\gamma}{\omega\epsilon} \frac{1}{r} \frac{\partial}{\partial r} (r J_r) \quad (1)$$

$$E_r = \frac{j}{\gamma} \frac{\partial E_z}{\partial r} - \frac{j\omega\mu}{\gamma^2} J_r \quad (2)$$

$$H_\theta = \frac{\omega\epsilon}{\gamma} E_r - \frac{j}{\gamma} J_r \quad (3)$$

TE WAVE

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial H_z}{\partial r} \right) - \gamma^2 H_z = -\frac{1}{r} \frac{\partial}{\partial r} (r J_\theta) \quad (4)$$

$$H_r = \frac{j}{\gamma} \left(\frac{\partial H_z}{\partial r} + J_\theta \right) \quad (5)$$

$$E_\theta = -\frac{\omega\mu}{\gamma} H_r = -\frac{j\omega\mu}{\gamma^2} \left(\frac{\partial H_z}{\partial r} + J_\theta \right) \quad (6)$$

Here (r, θ, z) are the polar cylindrical coordinates, ω the angular driving frequency, ϵ the dielectric constant and μ the permeability, of free space, in MKS units. The ac amplitudes of the electric and magnetic fields, and the convection-current density, respectively, are represented by the components of $\underline{E}$, $\underline{H}$, and $\underline{J}$. All ac quantities have been assumed to vary as $\exp j(\omega t - \gamma z)$.

When the assumption is made that the convection current density in the beam is of the same order of magnitude as the displacement current density, equations (2) and (6) reduce to the following:

$$E_r = \frac{j}{\gamma} \frac{\partial E_z}{\partial r} \quad (7)$$

$$E_\theta = -\frac{j\omega\mu}{\gamma^2} \frac{\partial H_z}{\partial r} \quad (8)$$

In order to evaluate the components of $\underline{J}$ in the beam, it is necessary to determine the velocity and charge distributions, first in the unmodulated, and then in the ac modulated beam.

The focusing of long cylindrical electron beams by axial magnetic fields of moderate strength has been fully described by Brillouin⁵ and Samuel⁶. This type of electron motion, called "Brillouin flow", can be established when a parallel electron beam abruptly enters a suitable magnetic field. The electrons thereupon acquire an angular velocity component which leads to a balance of radial forces in the beam.

The equations of motion of electrons in an axial magnetic field B_0 are as follows:

$$\ddot{r} - r\dot{\theta}^2 = \eta(\partial V_0/\partial r - r\dot{\theta}B_0) \quad (9)$$

$$r\ddot{\theta} + 2\dot{r}\dot{\theta} = \eta\dot{r}B_0 \quad (10)$$

$$\ddot{z} = \eta \cdot \partial V_0/\partial z \quad (11)$$

In these equations (r, θ, z) is the position of an electron at time t ; dots indicate differentiation with respect to t , following the electrons; $\eta = e/m$, where $-e$ is the electronic charge and m its mass; and V_0 is the potential describing the steady, axially symmetric electric field. Relativistic effects and the magnetic field resulting from electron motion have been neglected, as our interest is confined to beam velocities which are small compared to that of light.

It is readily verified that a solution of the above equation is:

$$\dot{r} = 0, \quad \dot{\theta} = \dot{\theta}_0 = \eta B_0/2, \quad \dot{z} = u_0 \quad (12)$$

$$\eta \partial V_0 / \partial r = r \dot{\theta}_0^2, \quad \partial V_0 / \partial z = 0 \quad (13)$$

Thus all the particles in the beam have the same angular velocity, equal to the Larmor angular frequency, and the same axial velocity u_0 . From Poisson's equation, we find the charge density:

$$\rho_0 = -2\epsilon \dot{\theta}_0^2 / \eta \quad (14)$$

It is convenient to introduce the angular plasma frequency ω_p , defined by:

$$\omega_p^2 = -\eta \rho_0 / \epsilon = 2\dot{\theta}_0^2 \quad (15)$$

In steady-state flow, an electron with initial position (r_0, θ_0, z_0) has the position $(r_0, \theta_0 + \dot{\theta}_0 t, z_0 + u_0 t)$ at time t . When the beam is modulated by a small ac signal, the electrons suffer small ac displacements from their steady-state trajectories. If we assume that the signal propagates along the axis of the beam as $\exp j(\omega t - \gamma z)$, we can write the perturbed electron coordinates in terms of the Lagrangian coordinates (r_0, θ_0, z_0) as follows:

$$r = r_0 + \tilde{r}(r_0) \cdot \exp j[\omega t - \gamma(z_0 + u_0 t)] \quad (16)$$

$$\theta = \theta_0 + \dot{\theta}_0 t + \tilde{\theta}(r_0) \cdot \exp j[\omega t - \gamma(z_0 + u_0 t)] \quad (17)$$

$$z = z_0 + u_0 t + \tilde{z}(r_0) \cdot \exp j[\omega t - \gamma(z_0 + u_0 t)] \quad (18)$$

where the tildes indicate ac amplitudes, and the dots indicate, as before, time differentiation at fixed r_0, θ_0, z_0 . Thus the dots are equivalent to multiplication by $j(\omega - \gamma u_0)$, when applied to ac quantities.

The equations of motion for the ac modulated beam differ from the steady-state equations (9) — (11), in that the particle coordinates are now given by (16) — (18), and there are ac fields present in addition to the dc fields $-\partial V_0 / \partial r$ and B_0 . As is usual in small-signal theory, only first-order ac quantities are retained in any equation. To this approxi-

mation, the ac fields can be evaluated at the unperturbed particle position.

Not all of the ac fields need to appear in the force equations, however. Reference to the field equations shows that the contributions of the ac magnetic fields to the force components are smaller than those due to the electric fields by a factor of the order of $(u_0/c)^2$ or smaller (where c is the velocity of light), and hence may be neglected. In addition, the force exerted by E_θ is of the same order as that due to H_r , and may be neglected too.

Omitting the factor $\exp j[\omega t - \gamma(z_0 + u_0 t)]$ for brevity from all ac terms, we can write the equations of motion as follows:

$$\ddot{r} - (r_0 + \tilde{r})(\dot{\theta}_0 + \dot{\tilde{\theta}})^2 = -\eta[-\partial V_0/\partial r + E_r + (r_0 + \tilde{r})(\theta_0 + \tilde{\theta})B_0] \quad (19)$$

$$(r_0 + \tilde{r})\ddot{\theta} + 2\dot{\tilde{r}}(\dot{\theta}_0 + \dot{\tilde{\theta}}) = \eta\dot{\tilde{r}}B_0 \quad (20)$$

$$\ddot{\tilde{z}} = -\eta E_z \quad (21)$$

These equations may be simplified with the aid of (12):

$$\eta \partial V_0/\partial r = (r_0 + \tilde{r})\dot{\theta}_0^2 \quad (22)$$

and by recalling that the dots may be replaced by multiplication by $j(\omega - \gamma u_0)$. We obtain, finally:

$$\tilde{r} = \eta E_r/(\omega - \gamma u_0)^2 \quad (23)$$

$$\tilde{\theta} = 0 \quad (24)$$

$$\tilde{z} = \eta E_z/(\omega - \gamma u_0)^2 \quad (25)$$

Although the foregoing equations deal with the dynamics of individual electrons, the assumption that the beam behaves like a smoothed-out "fluid" of charge, with a single velocity at each point, enables us to assign values of velocity and all other ac quantities, to *fixed positions* in space, (r, θ, z) . In these coordinates, the dc velocity is given by:

$$\underline{v}_0 = (0, r\dot{\theta}_0, u_0) \quad (26)$$

and the ac velocity by:

$$\begin{aligned} \underline{v} &= (\dot{\tilde{r}}, r\dot{\tilde{\theta}}, \dot{\tilde{z}}) \\ &= j(\omega - \gamma u_0)[(\tilde{r}, r\tilde{\theta}, \tilde{z})] \end{aligned} \quad (27)$$

Although the ac quantities are defined at r_0 , they may be taken to be the same at r , to a linear approximation.

The same result, (27), might have been obtained by stating the equations of motion in terms of Eulerian coordinates, in which the perturbed variables are the components of fluid velocity at any fixed point. In this procedure, the "material" or total time derivative would be used in the expressions for acceleration.

The ac space-charge density ρ is found with the aid of the continuity equation:

$$\frac{\partial}{\partial t} (\rho_0 + \rho) = -\text{div} [(\rho_0 + \rho)(\underline{v}_0 + \underline{v})] \quad (28)$$

$$\rho = \frac{j\rho_0}{\omega - \gamma u_0} \text{div } \underline{v} \quad (29)$$

From (23)–(25) and (27), the ac velocity may be written:

$$\underline{v} = \frac{j\eta}{\omega - \gamma u_0} (E_r, 0, E_z) \quad (30)$$

Combining these with Poisson's equation, we find:

$$\rho = -\frac{\eta\rho_0}{(\omega - \gamma u_0)^2} \text{div } \underline{E} = \frac{\omega_p^2}{(\omega - \gamma u_0)^2} \rho \quad (31)$$

There are two possible solutions to (31):

$$(\omega - \gamma u_0)^2 = \omega_p^2 \quad (32)$$

$$\rho = 0 \quad (33)$$

Solution (32) represents two longitudinal space-charge waves of arbitrary amplitude distribution, with plasma-frequency oscillations about the average beam velocity:

$$\gamma = \frac{\omega}{u_0} \pm \frac{\omega_p}{u_0} \quad (34)$$

The second solution, (33), however, permits us to evaluate the components of the ac convection current density $\underline{J}$, and thereby solve the field equations (1) — (8):

$$\underline{J} = \rho_0 \underline{v} + \rho \underline{v}_0 \quad (35)$$

$$J_r = \rho_0 v_r = \frac{\omega_p^2 \epsilon}{\gamma(\omega - \gamma u_0)} \frac{\partial E_z}{\partial r} \quad (36)$$

$$J_\theta = 0 \quad (37)$$

$$J_z = \rho_0 v_z = -\frac{j\omega_p^2 \epsilon}{\omega - \gamma u_0} E_z \quad (38)$$

The wave equations (1) and (4) for E_z and H_z now reduce to the following.

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial E_z}{\partial r} \right) - \gamma^2 E_z = 0 \quad (39)$$

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial H_z}{\partial r} \right) - \gamma^2 H_z = 0 \quad (40)$$

These equations have solutions for E_z and H_z , which are finite at $r = 0$, of the form $A \cdot I_0(\gamma r)$, where A is an arbitrary constant and I_0 the modified Bessel function of zero-th order.

It is not without interest to remark that the same pair of solutions, given by (32) and (39) — (40), has been found by L. R. Walker for a beam of arbitrary cross-section, with the same longitudinal velocity and space-charge density at every point, in the absence of any impressed dc magnetic field.

Due to the radial component of electron motion, the beam surface is rippled. For a steady-state radius b , this rippling can be expressed, in a linear approximation, by the perturbed radius:

$$r(b) = b + \tilde{r}(b) \exp j(\omega t - \gamma z) \quad (41)$$

The rippled beam is equivalent to a uniform cylindrical beam with an ac surface charge density $\rho_0 \tilde{r}$, or a surface current density whose components are:

$$G_z = \rho_0 \tilde{r} u_0 \quad (42)$$

$$G_\theta = \rho_0 \tilde{r} \dot{\theta}_0 b \quad (43)$$

The total ac convection current may be written in a form which applies equally well to the cylindrical beam with purely rectilinear flow:

$$\begin{aligned} I_c &= \int_0^b J_z 2\pi r dr + 2\pi b \rho_0 u_0 \tilde{r}(b) \\ &= -j\omega \epsilon \cdot R \cdot 2\pi b \cdot A \cdot I_1(\gamma b) / \gamma \\ &= -j\omega \epsilon R \int_0^b E_z 2\pi r dr \end{aligned} \quad (44)$$

where R is a beam propagation function which will prove convenient:

$$R = \frac{\omega_p^2}{(\omega - \gamma u_0)^2} = \frac{(\beta_p b)^2}{(\gamma b - \beta_e b)^2} \quad (45)$$

and

$$\beta_e = \omega / u_0, \quad \beta_p = \omega_p / u_0 \quad (46)$$

Thus we note that wave propagation along a cylindrical beam with Brillouin flow is accompanied by swelling and contracting of its boundary, with constant space-charge density, rather than by space-charge bunching. The second interesting result is that the dynamics and field equations for the focused beam are identical with those for a beam with zero dc magnetic field, except for the angular component of surface current density G_θ .

SPACE-CHARGE WAVES

We now consider the given beam, of radius b , in a concentric conducting tube of radius $a > b$. The boundary problem consists of matching the TM wave admittances inside and outside of the beam, at its boundary. (The TE fields are of no interest in the drift-tube problem, as they are not excited at the ends of the tube, and are not coupled to the TM fields.) Let I refer to the beam region $0 \leq r \leq b$, and II to the space between beam and conductor $b \leq r \leq a$. Then, at $r = b$,

$$\frac{H_\theta^I + G_z}{E_z^I} = \frac{H_\theta^{II}}{E_z^{II}} \quad (47)$$

The beam admittance on the left is evaluated with the aid of (3), (7), (36), and (42):

$$Y_e = \frac{j\omega\epsilon}{\gamma} (1 - R) \frac{I_1(\gamma b)}{I_0(\gamma b)} \quad (48)$$

In region II,

$$\begin{aligned} E_z &= B \cdot I_0(\gamma r) + C \cdot K_0(\gamma r) \\ H_\theta &= \frac{j\omega\epsilon}{\gamma} [B \cdot I_1(\gamma r) - C \cdot K_1(\gamma r)] \end{aligned}$$

where K_0 and K_1 are modified Bessel functions of the second kind. The wave admittance at $r = b$ in II is therefore:

$$Y_c = \frac{j\omega\epsilon}{\gamma} \left[\frac{I_1(\gamma b) - (C/B) \cdot K_1(\gamma b)}{I_0(\gamma b) + (C/B) \cdot K_0(\gamma b)} \right] \quad (49)$$

At $r = a$, $E_z^{II} = 0$ or:

$$C/B = -I_0(\gamma a)/K_0(\gamma a) \quad (50)$$

Equating beam and circuit admittances (48) and (49), we obtain:

$$R = \frac{-\frac{I_0(\gamma a)}{K_0(\gamma a)}}{\gamma b \cdot I_1(\gamma b) \cdot \left[I_0(\gamma b) - \frac{I_0(\gamma a)}{K_0(\gamma a)} \cdot K_0(\gamma b) \right]} \quad (51)$$

This equation must be solved simultaneously with each of the following:

$$\begin{aligned}\gamma_1 b &= \beta_e b + \beta_p b / \sqrt{R_1} \\ \gamma_2 b &= \beta_e b - \beta_p b / \sqrt{R_2}\end{aligned}\quad (52)$$

Thus, for a given beam and frequency, the solution consists of two unattenuated waves, one faster and the other slower than the beam velocity. The wavelength of the interference pattern is given by:

$$\lambda_s = \frac{4\pi}{\gamma_1 - \gamma_2} \quad (53)$$

For a cylindrical beam,

$$\beta_p b = 174 \sqrt{P} \quad (54)$$

where $P = I/V^{3/2}$ amps/(volts)^{3/2}, the perveance. In practice, P and hence $\beta_p b$ are usually so small that we can gain a fair estimate of λ_s by assuming $R_1 = R_2$:

$$\lambda_s \simeq \frac{2\pi \sqrt{R}}{\beta_p} \quad (55)$$

Fig. 1 shows the variation of $R^{1/2}$ with γb for several values of b/a . (The "intrinsic" solution (32) is included as a line at $R^{1/2} = 1$.) The ordinates of these curves are approximately proportional to the space-charge wavelength, and the abscissae to the frequency, as $\gamma \simeq \beta_e = \omega/u_0$ for small perveance.

Space-charge waves propagating along a cylindrical beam with rectilinear flow have been treated by Hahn⁸ and Ramo⁹. In Fig. 2, their computations have been reformulated in the same way as in Fig. 1, and compared with the results for Brillouin flow, for two values of b/a . The space-charge wavelength is always greater in Brillouin flow, for the principal pair of waves and the same b/a and γb .

HELIX PROBLEM

In place of the drift tube at radius a , we now have a helically conducting sheet of zero thickness and pitch angle ψ . In addition to I ($0 \leq r \leq b$) and II ($b \leq r \leq a$), we shall use III to identify fields in the region ($a \leq r < \infty$). The boundary conditions at $r = b$ are:

$$\begin{aligned}H_\theta^I + G_z - H_\theta^{II} &= 0 \\ E_z^I - E_z^{II} &= 0 \\ H_z^I - G_\theta - H_z^{II} &= 0 \\ E_\theta^I - E_\theta^{II} &= 0\end{aligned}\quad (56)$$

At $r = a$, the boundary conditions are:

$$\begin{aligned} E_z^{II} + E_\theta^{II} \cot \psi &= 0 \\ E_z^{III} + E_\theta^{III} \cot \psi &= 0 \\ E_z^{II} - E_z^{III} &= 0 \\ H_z^{II} + H_\theta^{II} \cot \psi - H_z^{III} - H_\theta^{III} \cot \psi &= 0 \end{aligned} \quad (57)$$

Inasmuch as $\cot \psi \sim \gamma/k$, the contribution of E_θ to the field at the helix can conceivably be comparable to that of E_z . The TE fields are coupled to the TM group, in addition, through the angular surface current G_θ , which depends on E_z . All 8 equations must therefore be solved simultaneously.

The procedure follows that of Chu and Jackson² for the field solution

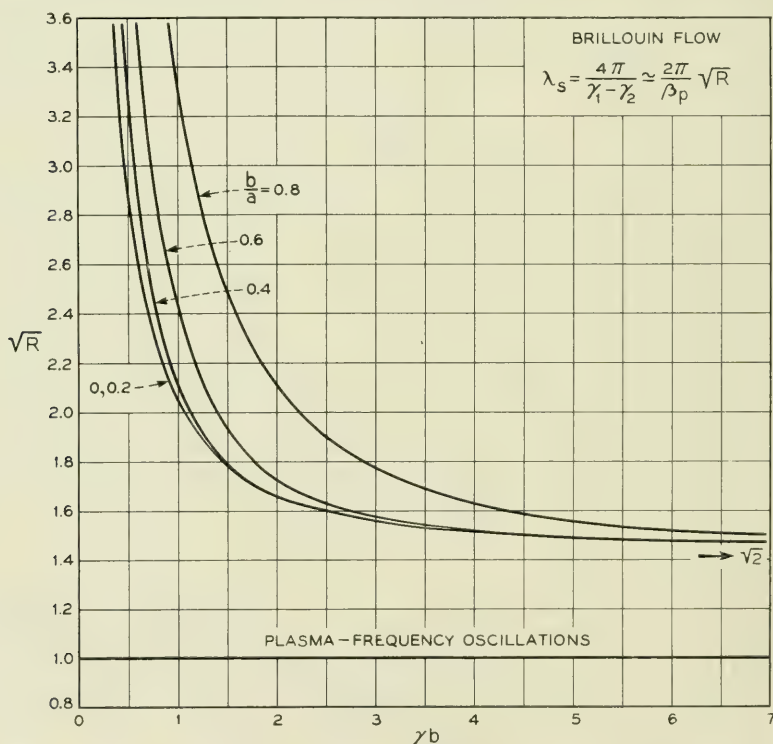


Fig. 1 — Space-charge wavelength λ_s for cylindrical beam with Brillouin flow, in a concentric drift tube. Here b and a are the beam and tube radii, respectively; $R^{1/2}$ is a dimensionless parameter; and the waves propagate as $\exp j(\omega t - \gamma z)$. To compute $\lambda_s \approx 2\pi R^{1/2}/\beta_p$, use $\beta_p b = 174 P^{1/2}$, where P is the beam perveance. The abscissae are approximately given by $\gamma \sim \beta e = \omega/u_0$. (Equations 52-55.)

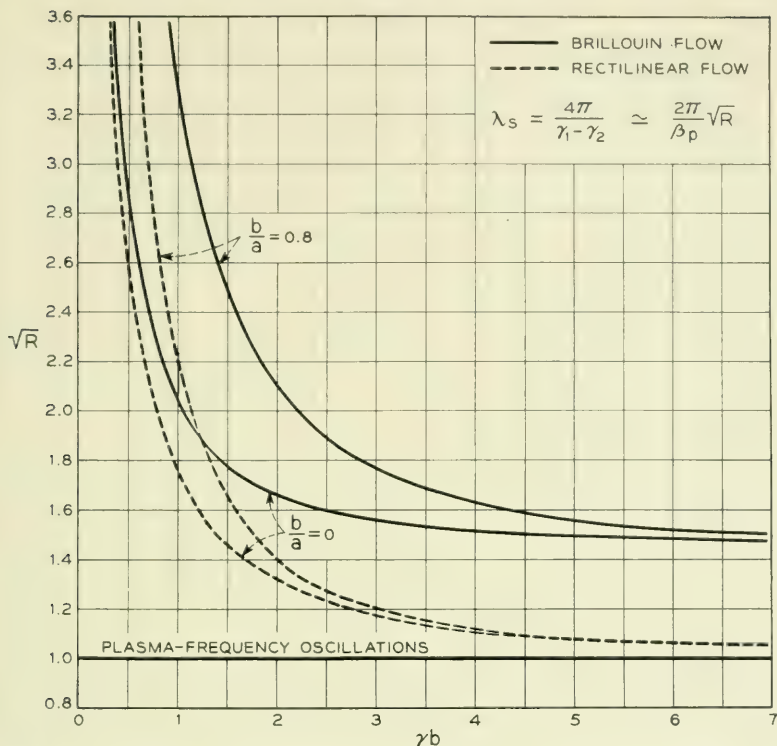


Fig. 2 — Comparison between space-charge wavelengths for cylindrical beams with Brillouin flow and those with rectilinear (confined) flow, respectively.

of the rectilinear beam. The 12 independent variables of (56–57) are reduced to 6 by expressing H_θ and E_θ in terms of E_z and H_z , respectively. The latter, however, require 2 arbitrary constants for a complete description in region II, making a total of 8 constants to be determined.

The eliminant of the 8 boundary-value equations can be written as a TM wave-admittance equation at the beam surface:

$$\frac{j\omega\epsilon}{\gamma} (1 - R) \frac{I_1(\gamma b)}{I_0(\gamma b)} = \frac{j\omega\epsilon}{\gamma} \frac{I_1(\gamma b) - \delta \cdot K_1(\gamma b)}{I_0(\gamma b) + \delta \cdot K_0(\gamma b)} \quad (58)$$

where

$$\delta = \frac{\delta_0 + RF}{1 - \frac{K_0(\gamma b)}{I_0(\gamma b)} RF} \quad (59)$$

$$\delta_0 = \frac{1}{K_0^2(\gamma a)} \left[\left(\frac{ka \cot \psi}{\gamma a} \right)^2 K_1(\gamma a) I_1(\gamma a) - K_0(\gamma a) I_0(\gamma a) \right] \quad (60)$$

$$F = (kb \cot \psi) \left(\frac{\dot{\theta}_0 b}{c} \right) \frac{I_1^2(\gamma b) K_1(\gamma a)}{K_0(\gamma a)} \quad (61)$$

In (61), c is the velocity of light.

The right side of (58), which is the admittance H_θ/E_z looking away from the beam surface, contains a term δ which depends on the helix geometry and the amplitude of the TE fields excited by the surface current G_θ . Thus, although the TE fields do not affect the electron paths, they are excited by the beam, and coupled to the TM fields at the helix. If G_θ were zero, δ would reduce to δ_0 , and the circuit admittance in (58) would then be the same as for a cylindrical beam with rectilinear flow. In (59), δ is expressed in terms of δ_0 and the product, RF , where R is the beam propagation function, and F a factor dependent on the magnetic field and the geometry.

This is the complete field solution of the problem. Equation (58) has four roots: two complex and two real propagation wave numbers, one of the latter representing a backward wave. In addition, there are two unattenuated space-charge waves, given by (34); or a total of 6 waves in all.

EQUIVALENT THIN-BEAM SOLUTION

Pierce¹ has expressed the admittance equation for an ideally thin beam, interacting with an arbitrary distributed circuit, as follows:

$$\frac{q}{E} = \frac{j\beta_e}{(j\beta_e - \Gamma)^2} \frac{I_0}{2V_0} = - \left(\Gamma^2 K \left[\frac{\Gamma_0}{\Gamma^2 - \Gamma_0^2} + \frac{2jQ}{\beta_e} \right] \right)^{-1} \quad (62)$$

where q = total convection current

E = longitudinal electric field

Γ = propagation constant = $\sqrt{-\gamma^2 - k^2}$

I_0 = dc beam current

V_0 = dc beam potential

Γ_0, K, Q = normal-mode circuit parameters.

For slow waves, $\Gamma \simeq j\gamma$. For moderate values of perveance, the accelerating voltage may be replaced by the beam potential at the axis:

$$u_0 \simeq \sqrt{2\eta V_0}$$

$$\frac{j\beta_e}{(j\beta_e - \Gamma)^2} \frac{I_0}{2V_0} \simeq j\omega \epsilon \pi b^2 R \quad (63)$$

Then, dividing both sides of (62) by $2\pi b$, we may rewrite it as a wave-admittance equation:

$$-\frac{j\omega\epsilon}{\gamma} \cdot \frac{\gamma b}{2} \cdot R = -\left(2\pi b \Gamma^2 K \left[\frac{\Gamma_0}{\Gamma^2 - \Gamma_0^2} + \frac{2jQ}{\beta_c} \right] \right)^{-1} \quad (64)$$

With the aid of (59), we can solve the admittance equation (58) for R , and re-write it as follows:

$$Y_e = Y_B \quad (65)$$

with

$$Y_e = -\frac{j\omega\epsilon}{\gamma} \cdot \frac{\gamma b}{2} \cdot R \quad (66)$$

$$Y_B = \frac{j\omega\epsilon}{\gamma} \cdot \frac{\gamma b}{2} \cdot \frac{I_0(\gamma b)}{I_1(\gamma b)} \left(-\frac{\delta_0}{\gamma b \cdot I_0(\gamma b) [I_0(\gamma b) + \epsilon_0 \cdot K_0(\gamma b)] - \frac{FI_0(\gamma b)}{I_1(\gamma b)}} \right) \quad (67)$$

The solid-cylindrical Brillouin beam in a helix is thus equivalent to a thin beam whose circuit admittance is Y_B . By equating Y_B to the right side of (64), we can evaluate the normal-mode parameters for this admittance, and thereby use all the results of previous thin-beam calculations.^{1,7} The equivalence of the two circuit expressions, however, requires that we replace the transcendental expression (67) by an algebraic one, with no more than three arbitrary constants. This can be done very effectively, in the region of interest, by means of the approximation:²

$$Y_B \simeq -(\gamma_p - \gamma_0) \left(\frac{\partial Y_B}{\partial \gamma} \right)_{\gamma=\gamma_0} \cdot \frac{\gamma - \gamma_0}{\gamma - \gamma_p} \quad (68)$$

in which γ_0 and γ_p are the zero and pole, respectively, of Y_B :

$$\delta_0(\gamma_0) = 0$$

$$\delta_0(\gamma_p) = \left(-\frac{I_0(\gamma b)}{K_0(\gamma b)} + \frac{F}{\gamma b \cdot I_1(\gamma b) \cdot K_0(\gamma b)} \right)_{\gamma=\gamma_p} \quad (70)$$

If we were to neglect the term containing F in (70), the error in the magnitude of $\delta_0(\gamma_p)$ would be measured by:

$$\frac{F}{\gamma b \cdot I_1(\gamma b) \cdot I_0(\gamma b)} \equiv (kb \cot \psi) \left(\frac{\partial_0 b}{c} \right) \frac{I_1(\gamma b) \cdot K_1(\gamma a)}{I_0(\gamma b) \cdot K_0(\gamma a)} \quad (71)$$

In most low-power traveling-wave tubes, the first factor in parentheses is usually less than 3; the second factor less than 0.01; and the last factor always less than unity. The error in evaluating γ_p , moreover, is less than

this product, for the slope of the curve δ_0 versus γb increases with γb . Putting $F = 0$, therefore, leads at most to a very slight error in γ_p and

$$\left(\frac{\partial Y_B}{\partial \gamma} \right)_{\gamma=\gamma_0}$$

Outside of the region (γ_0, γ_p) , δ_0 grows large rapidly, and the expression for Y_B is hardly affected at all by this assumption.

Physically, the negligible role played by F in the admittance equation means that $\delta \simeq \delta_0$, i.e., the TM helix admittance is not appreciably perturbed by the TE fields excited by G_θ .

With $F = 0$, (67) may be re-written:

$$Y_B = \frac{\gamma b}{2} \cdot \frac{I_0(\gamma b)}{I_1(\gamma b)} \cdot Y_H \quad (72)$$

$$Y_H = \frac{j\omega\epsilon}{\gamma} \left[\frac{I_1(\gamma b) - \delta_0 \cdot K_1(\gamma b)}{I_0(\gamma b) + \delta_0 \cdot K_0(\gamma b)} - \frac{I_1(\gamma b)}{I_0(\gamma b)} \right] \quad (73)$$

Here Y_H is the helix admittance seen by a thin cylindrical hollow beam, with rectilinear electron flow. As in the case of Y_B , it may be replaced in the vicinity of (γ_0, γ_p) , by the approximation:

$$Y_H \simeq -(\gamma_p - \gamma_0) \left(\frac{\partial Y_H}{\partial \gamma} \right)_{\gamma=\gamma_0} \frac{\gamma - \gamma_0}{\gamma - \gamma_p} \quad (74)$$

Fletcher⁴ has evaluated the normal-mode parameters for Y_H as follows:

$$\Gamma_0^2 = -\gamma_0^2 - k^2 \quad (75)$$

$$Q_H \frac{\gamma_0}{\beta_e} \left[1 + \frac{k^2}{\gamma_0^2} \right]^{-1/2} = \frac{1}{2} \frac{\gamma_0^2}{\gamma_p^2 - \gamma_0^2} \quad (76)$$

$$\frac{1}{K_H} = -j\pi b \gamma_0^2 \left[1 + \frac{k^2}{\gamma_0^2} \right]^{3/2} \left(\frac{\partial Y_H}{\partial \gamma} \right)_{\gamma=\gamma_0} \quad (77)$$

We have used the subscript H to refer the parameters to the hollow beam, and will use the subscript B to refer to the solid-cylindrical Brillouin beam.

As Y_B and Y_H have the same zero and pole, they have the same natural propagation constant Γ_0 , and the same space-charge parameter Q :

$$Q_B = Q_H \quad (78)$$

This quantity can be found plotted in Fig. 1 of Reference 4, or in Fig. A6.1 of Reference 1.

From (68), (72), and (74), we find:

$$\left(\frac{Y_B}{Y_H} \right)_{\gamma=\gamma_0} = \left(\frac{\partial Y_B / \partial \gamma}{\partial Y_H / \partial \gamma} \right)_{\gamma=\gamma_0} = \left[\frac{\gamma b}{2} \frac{I_0(\gamma b)}{I_1(\gamma b)} \right]_{\gamma=\gamma_0} \quad (79)$$

The impedance parameters for the two beams are therefore related to each other by:

$$K_B = K_H \left[\frac{2}{\gamma b} \frac{I_1(\gamma b)}{I_0(\gamma b)} \right]_{\gamma=\gamma_0} \quad (80)$$

Both Pierce¹⁰ and Fletcher⁴ have found the impedance parameter of the hollow beam to be related to that of a thin beam along the axis of a helix, K_T , as follows:

$$K_H = K_T [I_0^2(\gamma b)]_{\gamma=\gamma_0} \quad (81)$$

$$\text{The gain parameter } C \text{ is defined by: } C^3 = (2K)(I_0/8V_0) \quad (82)$$

Thus, for given I_0 and V_0 , the factor by which the gain parameter of a thin beam should be multiplied to give that of a hollow beam, is:

$$(K_H/K_T)^{1/3} = [I_0^{2/3}(\gamma b)]_{\gamma=\gamma_0} \quad (83)$$

This "impedance reduction factor" can similarly be evaluated for the finite cylindrical beam with Brillouin flow:

$$(K_B/K_T)^{1/3} = \left[\frac{2}{\gamma b} I_1(\gamma b) \cdot I_0(\gamma b) \right]_{\gamma=\gamma_0}^{1/3} \quad (84)$$

Cutler,⁷ who calls this quantity F_2 , has described how it and the parameter Q can be used to compute the gain of traveling-wave tubes. The procedure depends upon the evaluation of C and QC . The expression for C_B , in Cutler's notation, is:

$$C_B \simeq (K_2/K)^{1/3} F_1 F_2 (I_0/8V_0)^{1/3} \quad (85)$$

Here K_2/K is a factor, of the order of 0.5, which corrects the impedance of the ideal sheath helix for the physical dimensions and support elements of the actual helix. It is best found by measurement. The factor F_1 is plotted in Fig. 3.4 of Reference 1, and obeys the empirical relation:

$$F_1(\gamma a) = 7.154 \exp(-0.6664 \gamma a) \quad (86)$$

Finally, the factor F_2 is the impedance reduction factor (84), which is plotted in Fig. 3 of this paper for various ratios of the radii, b/a .

It is of interest to compare the relative gain of beams with rectilinear and with Brillouin flow, respectively. Pierce¹⁰ has computed a first approximation to the impedance reduction factor for the solid-cylindrical

beam with rectilinear flow, by averaging E_z^2 over the beam area (with E_z for the empty helix):

$$(K_s/K_T)^{1/3} \sim [I_0^2(\gamma b) - I_1^2(\gamma b)]_{\gamma=\gamma_0}^{1/3} \quad (87)$$

Fletcher⁴ has improved upon this calculation by replacing the solid beam with a thin hollow beam of different radius and dc current. This has the same electronic admittance Y_e and derivative $dY_e/d\gamma$ when $R = 1$.

The impedance reduction factors for the three types of beams have been plotted in Fig. 4, using a typical value of b/a . For the same I_0 , V_0 , and b/a , the gain parameters C are found to be greatest for the hollow beam, and least for the solid rectilinear beam.

The high gain of the hollow beam is due to its concentration in the region of greatest field strength. The greater gain of the beam with Brillouin flow, relative to that of a similar beam with confined flow, is probably due to transverse electron motion, in two ways:¹¹

(1) causing electrons to interact with the transverse as well as longitudinal fields; and

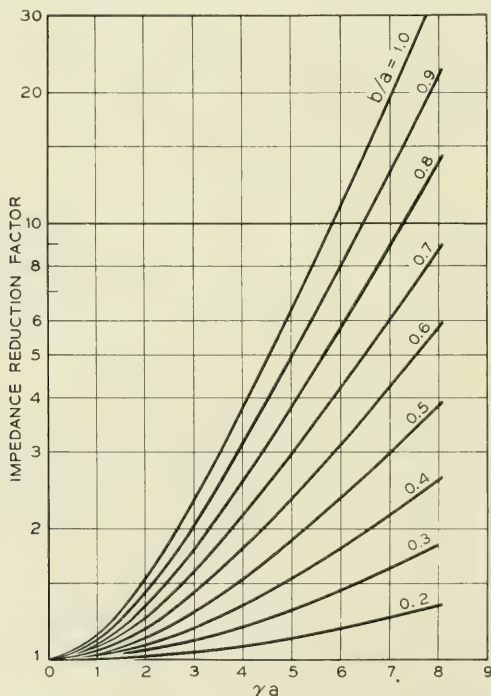


Fig. 3 — The factor $(K_B/K_T)^{1/3}$, or F_2 , by which the gain parameter C_T for a thin beam should be multiplied to give C_B , the gain parameter for a cylindrical beam with Brillouin flow, of the same current and voltage. Computation of C_B using this factor is described in text following equation (85).

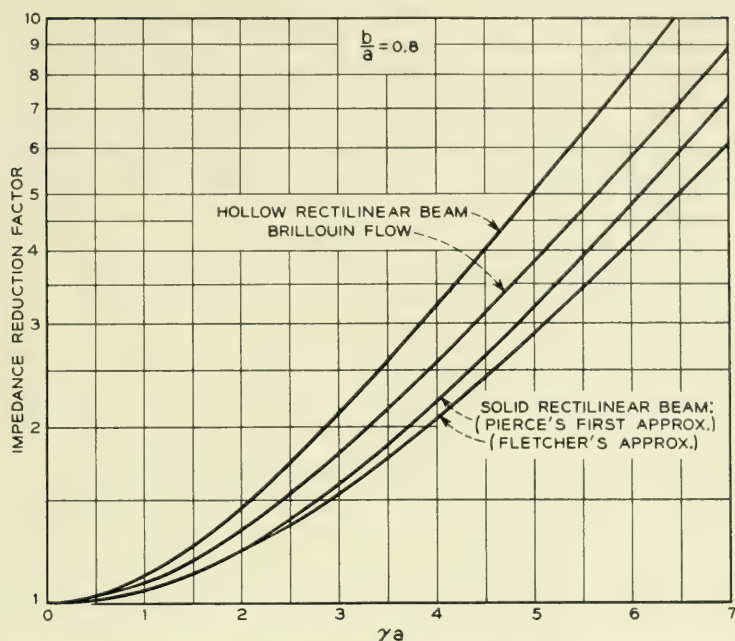


Fig. 4 — Comparative values of impedance reduction factor for several kinds of beams, of the same relative radii b/a .

(2) causing electrons to move preferentially into regions of retarding longitudinal fields, a process analogous to bunching.

CONCLUSIONS

Field solutions have been presented for the magnetically-focused cylindrical beam, when modulated by a small ac signal. Two types of beam enclosure have been treated: the concentric drift tube and the ideally thin (sheath) helix.

There are two pairs of unattenuated space-charge waves in the drift-tube: one with arbitrary amplitude distribution, and another pair which is coupled to the external field (Fig. 1). The space-charge wavelength of the latter pair is greater than that of space-charge waves in a similar beam with rectilinear flow (Fig. 2).

The solution of the helix problem consists of the aforementioned two space-charge waves with arbitrary amplitude, as well as the usual four waves of traveling-wave tube theory, or six waves in all. In order to compute the gain constant of the growing wave, the field solution has been re-written as the admittance equation of a thin beam in an artificial

circuit. By means of two approximations, the normal-mode parameters of this circuit have been evaluated.

The first approximation amounts to neglecting the TE fields coupled to the TM wave, and is valid for most low-power traveling-wave tubes. The second approximation consists of replacing the circuit admittance function by an algebraic expression with the same zero and pole, and the same slope at the zero. Although thin-beam theory predicts small deviations of complex roots (of the admittance equation) from the natural propagation wave number, it is difficult to judge whether any such roots might occur outside of the region in which this approximation holds, for the finite beam.

The space-charge parameter Q_B is found to be the same as for a thin hollow beam with rectilinear flow (Fig. 1 of Reference 4, or Fig. A6.1 of Reference 1). The gain parameter C_B can be computed from Equation (85), Fig. 3.4 of Reference 1, and Fig. 3 of this paper. The gain of the cylindrical beam with Brillouin flow is found to be greater than that of a similar cylindrical beam with rectilinear flow, presumably because of transverse electron motion in the former. Its gain, however, is less than that of a thin hollow cylindrical beam with rectilinear flow, for the same radius, current, and voltage (Fig. 4).

ACKNOWLEDGEMENTS

The writers are indebted to J. R. Pierce for suggesting the general approach used in this paper, to H. Suhl and C. F. Quate for their helpful criticisms, and to Mrs. C. Lambert for computing the graph material.

REFERENCES

1. J. R. Pierce, *Traveling-Wave Tubes*, D. Van Nostrand, N. Y., Chapters VII, VIII, (1950).
2. L. J. Chu and J. D. Jackson, Field Theory of Traveling-Wave Tubes, Proc. I.R.E., **36**, p. 853, 1948.
3. E. H. Rydbeck, Theory of the Traveling-Wave Tube, Ericsson Tech. **46**, p. 3, 1948.
4. R. C. Fletcher, Helix Parameters Used in Traveling-Wave Tube Theory, Proc. I.R.E., **38**, p. 413, 1950.
5. L. Brillouin, A Theorem of Larmor and Its Importance for Electrons in Magnetic Fields, Phys. Rev. **67**, p. 260, 1945.
6. A. L. Samuel, On the Theory of Axially Symmetric Electron Beams in an Axial Magnetic Field, Proc. I.R.E., **37**, p. 1252, 1949.
7. C. C. Cutler, The Calculation of Traveling-Wave Tube Gain, Proc. I.R.E. **39**, p. 914, 1951.
8. W. C. Hahn, Small Signal Theory of Velocity-Modulated Electron Beams, G. E. Rev., **42**, p. 258, 1939.
9. S. Ramo, Space Charge and Field Waves in an Electron Beam, Phys. Rev., **56**, p. 276, 1939.
10. J. R. Pierce, *Traveling-Wave Tubes*, D. Van Nostrand, N. Y., App. II.
11. J. R. Pierce, *Traveling-Wave Tubes*, D. Van Nostrand, N. Y., Chapt. XIII.

Diffraction of Plane Radio Waves by a Parabolic Cylinder

Calculation of Shadows Behind Hills

By S. O. RICE

(Manuscript received April 3, 1953)

Expressions are given for the diffraction field far behind, and the surface currents on, a parabolic cylinder. Approximate values for the field strength and current density are given when the radius of curvature of the cylinder is large compared to a wavelength. The formulas may have value in predicting the shadows that are cast by hills in microwave propagation. The idea of representing hills by knife-edges has been used successfully by a number of investigators. The theory of the parabolic cylinder indicates that such a representation is valid even for gently rounded hills when the angle of diffraction is small. On the other hand, when the angle of diffraction is so large that the knife-edge calculations do not apply, the results presented here may be used.

1. INTRODUCTION

A number of investigators have studied the effect of hills on the propagation of short radio waves. Experiment has shown that the field far behind a hill may be computed, to a reasonable degree of accuracy, by assuming that the hill acts like a knife-edge (half-plane)¹. The question naturally arises as to the conditions under which such an assumption is permissible. Here we attempt to throw some light on this question by taking the hill to be a parabolic cylinder.

Our results indicate that, for small angles of diffraction, even gently curved hills act like knife-edges. However, for larger angles corresponding to points deep in the shadow or to points high in the illuminated region, it may be necessary to use the more exact formulas which take the curvature of the hill into account.

¹ See, for example, Ultra-Short-Wave Propagation, J. C. Schelleng, C. R. Burrows and E. B. Ferrell, Proc. I.R.E., **21**, pp. 423-463, 1933.

As an application of our results, a brief study is made of the hypothesis that very short radio waves are transmitted far beyond the horizon by successive diffractions over hills or ridges. The ridges are assumed to be of equal height, to be 40 miles apart, and to have a radius of curvature of 100 meters at their crests. At the frequencies considered (30 and 300 mc) and at the small angles of diffraction required to go from one crest to the next, the ridges act like knife-edges.

At 30 mc and a distance of 280 miles the calculated field is in fair agreement with the observed field². At 300 mc and at the same distance the calculated field is about 50 db below the observed field. This suggests that the 30 mc long distance propagation may possibly be explained by successive diffractions. The discrepancy at 300 mc may be due to any one of a number of reasons. For example, it may be due to the effect of the non-uniformity of the atmosphere, or to the roughness of our approximations (for one thing, we neglect reflections from the ground between the ridges).

The first theoretical work on the diffraction of plane electromagnetic waves by a parabolic cylinder apparently was done by P. S. Epstein.³ His work makes use of a series of parabolic cylinder functions. When the cylinder is large many terms are required for computation. By "large" we mean that the radius of curvature at the vertex of the cylinder is large compared to the wavelength of the radio wave.

An entirely different approach was used by V. Fock^{4,5} in 1946. In the first paper Fock sketches the derivation of an integral for the current density on a large paraboloid of revolution. In the second paper he re-derives this integral by considering the form assumed by the field equations when a plane wave strikes a gently curved conductor at grazing incidence. His result gives the change in current density on a large and highly conducting parabolic cylinder as we go from the illuminated region into the shadow.

In 1950 K. Artmann⁶ examined the diffraction field far behind a large circular cylinder. He showed that, for small angles of diffraction, the

² A summary of experimental data is given by K. Bullington, Radio Transmission beyond the Horizon in the 40-400 Megacycle Band, Proc. I.R.E., **41**, pp. 132-135, 1953.

³ Dissertation, Munich (1914). A more accessible account of this work is given in the *Encyklopädie der Math. Wiss.* 5, Part 3 (1909-1926) Phys. p. 511. See also H. Bateman, *Partial Differential Equations of Math. Phys.*, (Cambridge Univ. Press 1932) p. 488.

⁴ The Distribution of Currents Induced by a Plane Wave on the Surface of a Conductor, J. Phys. (U.S.S.R.), **10**, pp. 130-136, 1946.

⁵ The Field of a Plane Wave Near the Surface of a Conducting Body, J. Phys. (U.S.S.R.), **10**, pp. 399-409, 1946.

⁶ Beugung polarisierten Lichtes an Blenden endlicher Dicke im Gebiet der Schattengrenze, Zeitschr. für Phys. **127**, pp. 468-494, 1950.

diffraction pattern is shifted by an amount proportional to the $\frac{1}{3}$ power of the radius of curvature. Whether the shift is towards the shadow or in the opposite direction depends upon the polarization of the incident wave.

In this paper we derive some of the results given by Fock and Artmann by starting with Epstein's series. In addition we investigate the diffracted field at a great distance behind the cylinder. The cylinder is assumed to be a perfect conductor in all of our work except for a few equations given near the ends of Sections 4, 6, and 7. The procedure is similar to that used in the smooth-earth theory.⁷ The series is converted into an integral and then the path of integration is deformed so as to gain as much simplification as possible. As might be expected, the results for a large parabolic cylinder are similar in some respects to those for a smooth earth. Much of the work requires a knowledge of the behavior of parabolic cylinder functions of large complex order. Although several studies of this behavior have been published, the results are not in the form required. For this reason, and for the sake of completeness, several sections dealing with the properties of parabolic cylinder functions have been included.

Incidentally, W. Magnus⁸ has studied the field produced by a line source located at the focus of a parabolic cylinder. However, his problem is somewhat different from the one with which we are concerned.

I am grateful to Prof. Erdélyi of the California Institute of Technology and to my colleagues at Bell Telephone Laboratories for helpful discussions and references which resulted in a number of improvements throughout the paper. I am also indebted to Miss Marian Darville for performing the rather laborious computations upon which the various curves and tables are based.

2. DISCUSSION OF RESULTS

Various expressions are given later for the electromagnetic field in terms of parabolic cylinder functions. In this section we shall confine a good share of our attention to the case in which the cylinder is very large compared to a wavelength so that the cylinder functions may be approximated by Airy integrals. As in the remainder of the paper, we shall be concerned chiefly with the field behind the cylinder and the current density on the cylinder.

⁷ By "Smooth-earth theory" we mean the formulas for the field produced by a dipole near a large sphere. A complete discussion of the theory is given in the book by H. Bremmer, *Terrestrial Radio Waves* (Elsevier, 1949).

⁸ Zur Theorie des zylindrisch-parabolischen Spiegels, *Zeitschr. für Physik*, **118**, pp. 343-356, 1941.

It turns out that the results for the parabolic cylinder are closely related to those obtained by Sommerfeld⁹ for the diffraction of a plane wave by a perfectly conducting half-plane. In fact, the two fields are surprisingly similar in the region of the shadow boundary. More precisely, the fields are similar for values of the angle ψ , defined in Fig. 2.3, such that (roughly)

$$|\psi \text{ in radians}| < \frac{1}{4} \left[\frac{\text{wavelength}}{\text{radius of curvature of cylinder at its crest}} \right]^{1/3},$$

where the coefficient $\frac{1}{4}$ has been selected somewhat arbitrarily. For larger values of $|\psi|$ the difference between the fields becomes pronounced. As we go deeper and deeper into the shadow, i.e. as ψ becomes more and more negative, the field behind a cylinder ultimately decreases exponentially with ψ . On the other hand, the field behind a half-plane decreases roughly as $1/|\psi|$. Since the exponential function decreases more rapidly than does $1/|\psi|$, the shadow behind a hill is darker than the one behind a half-plane. High in the illuminated region the field consists of the incident wave plus the wave reflected from the illuminated portion of the cylinder. For the half-plane this reflected wave is negligibly small until ψ reaches 180° .

First we shall review the situation pictured in Fig. 2.1. An incident wave comes in from the left and strikes a perfectly conducting vertical half-plane which casts a shadow as shown. The electric and magnetic intensities are proportional to $\exp(i\omega t)$ where t is the time and ω is the radian frequency. The unit of length is chosen so that λ , the wavelength, is equal to 2π . This is done in order to simplify the expressions we have to deal with. For example, a plane wave of unit amplitude traveling in the positive x direction, as shown in Fig. 2.1, is represented by $\exp(-ix)$.

Sommerfeld's exact expressions, for the special case of horizontal incidence shown in Fig. 2.1, may be written as

$$(\text{hp}) \quad E = (e^{-ix} + S_1) + S_2(0), \quad (2.1)$$

$$(\text{vp}) \quad H = (e^{-ix} + S_1) + S_3(0), \quad (2.2)$$

where (2.1) holds when the electric intensity E is parallel to the edge, and (2.2) when the magnetic intensity H is parallel to the edge. From

⁹ Math. Annalen, **47**, p. 317, 1896. Sommerfeld's results have been described in a number of texts on optics. The book, *Huygens' Principle* by Baker and Copson (Oxford 2nd edition, 1950) deals with this and many similar problems. See also Chap. 20 of Frank-von Mises, *Die Differential- und Integralgleichungen der Mechanik und Physik*, 2nd edition, Braunschweig: F. Vieweg and Sohn (1935).

the analogy with the radio case, these two polarizations will be termed "horizontal polarization" (hp) and "vertical polarization" (vp), respectively. The incident plane wave for hp is assumed to have an E of unit amplitude. This is indicated by the $\exp(-ix)$ in (2.1). For vp the incident wave is assumed to have an H of unit amplitude. The S 's are defined by the Fresnel integrals

$$\begin{aligned} e^{-ix} + S_1 &= (i/\pi)^{1/2} e^{-ix} \int_{-\infty}^{t_1} e^{-it^2} dt, \\ S_1 &= -(i/\pi)^{1/2} e^{-ix} \int_{t_1}^{\infty} e^{-it^2} dt, \\ S_2(0) &= -S_3(0) = -(i/\pi)^{1/2} e^{+ix} \int_{t_2}^{\infty} e^{-it^2} dt, \end{aligned} \quad (2.3)$$

$$t_1 = (2r)^{1/2} \sin \frac{\varphi}{2}, \quad t_2 = (2r)^{1/2} \cos \frac{\varphi}{2}, \quad i^{1/2} = \exp(i\pi/4),$$

where (r, φ) are the polar coordinates shown in Fig. 2.1 [S_1 and $S_2(0) = -S_3(0)$ for an arbitrary angle of incidence are given by Equations (5.3), (5.6), and (5.20) of Section 5]. We use the notation $S_2(0)$, $S_3(0)$ to indicate that these functions are special cases of the functions $S_2(h)$, $S_3(h)$ which appear in the analysis for the parabolic cylinder.

The principal part of the field far to the right of the half-plane, where x is positive, is given by $\exp(-ix) + S_1$ whose absolute value is plotted in Fig. 2.2. The function S_1 almost cancels the incident wave in the shadow, and then drops down to small values outside the shadow. The function $S_2(0)$ is always small in the region we shall consider. It becomes large only when φ exceeds π . It then corresponds to the wave reflected from the front (left-hand side) of the half-plane.

When we are far enough away from the shadow boundary to make

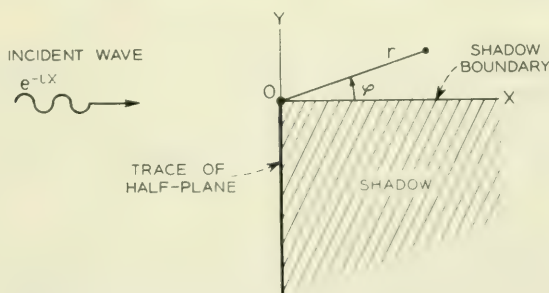


Fig. 2.1 — Diffraction of a plane wave by a half-plane.

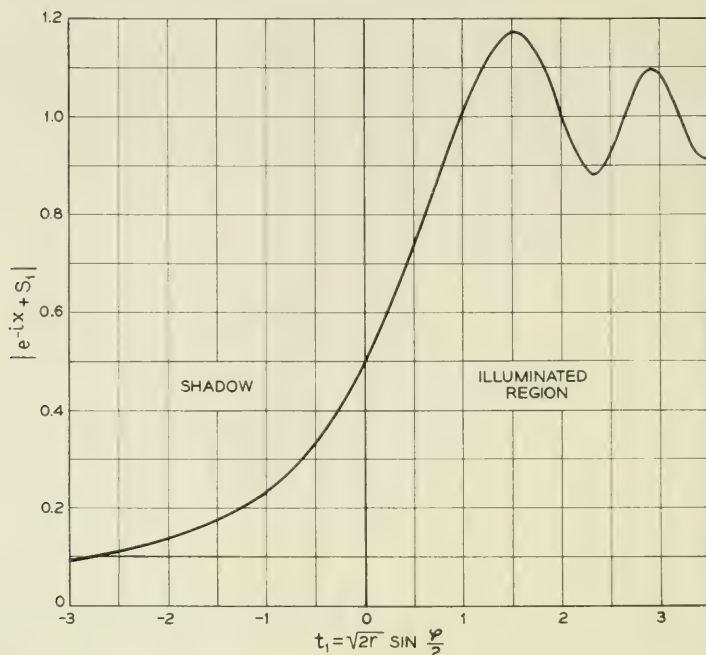


Fig. 2.2 — The approximate value $|e^{-ix} + S_1|$ of $|E|$ for hp and of $|H|$ for vp at a great distance r behind a half-plane. Here, as in all of our work, the wavelength λ is 2π . For an arbitrary wavelength replace r by $2\pi r/\lambda$, etc.

$\varphi^2 r \gg 1$, $\exp(-ix) + S_1$ has the asymptotic expressions [see Equations (7.7) and (7.8)]

$$e^{-ix} + S_1 \sim \begin{cases} c(r)/2 \sin \frac{\varphi}{2}, & \varphi < 0 \text{ (shadow)} \\ e^{-ix} + c(r)/2 \sin \frac{\varphi}{2}, & \varphi > 0 \end{cases} \quad (2.4)$$

$$c(r) = i^{3/2} (2\pi r)^{-1/2} e^{-ir}. \quad (2.5)$$

These expressions lead us to picture the field to the right of the half-plane as the sum of the incident wave and a wave, $c(r)/\varphi$, spreading out from the diffracting edge¹⁰ (for the small φ 's of interest, $2 \sin \varphi/2 \approx \varphi$ even though $\varphi^2 r \gg 1$). In the illuminated region these two waves interfere with each other to give the oscillations around unity shown in Fig.

¹⁰ See, for example, R. W. Wood, *Physical Optics*, 3rd edition, p. 220 (MacMillan, 1935). Curves of equal phase and amplitude have been computed and plotted by W. Braunbeck and G. Laukien, *Einzelheiten zur Halbebenen-Beugung*, *Optik*, **9**, pp. 174-179, 1952.

2.2. In the shadow only the edge wave is present and there is no interference. S_1 is not the edge wave in the shadow.

The edge does not radiate uniformly in all directions. The φ in the denominator of $c(r)/\varphi$ indicates that the edge sends out its strongest wave in the direction of the shadow boundary where $\varphi = 0$. This accounts for the decreasing size of the oscillations in Fig. 2.2 as φ becomes more and more positive. It likewise accounts for the steady decrease as φ becomes more and more negative.

That $S_2(0)$ is small in comparison with $\exp(-ix) + S_1$ follows from

$$S_2(0) \sim c(r)/2 \cos \frac{\varphi}{2} \quad (2.6)$$

and the fact that this is small compared to the $c(r)/2 \sin(\varphi/2)$ in (2.4) when φ is small.

So far, we have been discussing a special case of Sommerfeld's results. Now we turn to the case of the perfectly conducting parabolic cylinder shown in Fig. 2.3. Here, as in Fig. 2.1, the incident plane wave $\exp(-ix)$ comes in from the left. The fields for the two kinds of polarization are given by

$$(hp) \quad E = (e^{-ix} + S_1) + S_2(h), \quad (2.7)$$

$$(vp) \quad H = (e^{-ix} + S_1) + S_3(h), \quad (2.8)$$

where, just as in the half-plane case, the fundamental vectors E and H are perpendicular to the plane of Fig. 2.3 and the incident waves are of unit amplitude.

The $[\exp(-ix) + S_1]$ in (2.7) and (2.8) is exactly the same Fresnel integral (2.3) as for the half-plane. $S_2(h)$ is a rather complicated integral

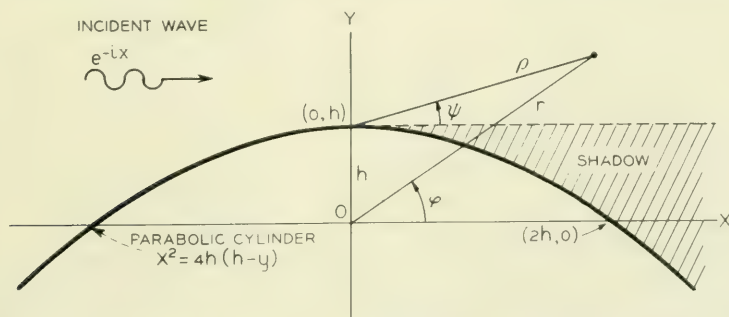


Fig. 2.3 - Coordinates used in the discussion of the parabolic cylinder. The coordinates such as $(0, h)$ refer to (x, y) . The origin 0 of coordinates is at the focus of the parabola and h is the height of the vertex or crest, above the origin.

(obtained by setting the angle of incidence θ equal to $\pi/2$ in equation (5.4)) involving parabolic cylinder functions. When $h = 0$ the parabolic cylinder reduces to a half-plane and $S_2(h)$ reduces to the value $S_2(0)$ appearing in (2.3). The symbol $S_3(h)$ represents an integral much like $S_2(h)$ except that it contains derivatives of the parabolic cylinder functions. As we might expect, $S_2(h)$ and $S_3(h)$ behave in much the same way as does $S_2(0)$. In particular, they are small compared to $\exp(-ix) + S_1$ at the shadow boundary, and their asymptotic expressions analogous to (2.6) hold as φ and ψ pass through zero.

$S_2(h)$ and $S_3(h)$ have been put in a form suitable for computation in two cases, (1) when $h = 0$, which is the half-plane case already discussed, and (2) when h and r/h^2 are very large. In the second case it is convenient to introduce new polar coordinates (ρ, ψ) with their origin at the crest of the parabola as shown in Fig. 2.3. In these coordinates a circular cylindrical wave spreading out from the crest is asymptotically proportional to

$$c(\rho) = i^{3/2} (2\pi\rho)^{-1/2} e^{-i\rho}. \quad (2.9)$$

In Section 8 it is shown that

$$E \approx [e^{-ix} + S_1]_\rho - \frac{c(\rho)}{\psi} + c(\rho)h^{1/3} \left[\Psi(\tau) + \frac{1}{\tau} \right] \exp(i\tau^3/3), \quad (2.10)$$

$$\tau = h^{1/3}\psi$$

is an approximation which gives the field (for horizontal polarization) in the region of the shadow boundary far behind a large cylinder. Our τ is an approximation to the g used there. Here the subscript ρ on $[\exp(-ix) + S_1]_\rho$ means that the quantity within the brackets is to be computed as though it corresponded to a half-plane with its edge at the crest of the parabolic cylinder, so that ρ, ψ are to be used in (2.3) instead of r, φ . Also,

$$\Psi(\tau) = 2i^{-1/3} \int_0^\infty \frac{Ai(u) \exp(i^{-1/3}u\tau) du}{Ai(u) - iBi(u)} + 2i \int_0^\infty \frac{Ai(u) \exp(iu\tau) du}{Ai(u) + iBi(u)} \quad (2.11)$$

where $Ai(u)$ and $Bi(u)$ are Airy integrals defined by equations (13.12) and (13.16), and tabulated in reference.¹¹ In this paper we find it convenient to use the Airy integrals instead of the related Bessel functions

¹¹ The Airy Integral, Brit. Asso. Math. Tables, Part — Vol. B (Cambridge, 1946).

of order $1/3$. As in (2.3), the fractional powers of i are made precise by taking $i \equiv \exp(i\pi/2)$.

Three kinds of approximations have been made in the derivation of (2.10), namely those associated with the assumptions (1) that the angle ψ is small, (2) that h is large, and (3) that r/h^2 is large. The terms in $1/\psi$ and $1/\tau$ do not cause trouble at $\psi = 0$ because their infinities cancel each other.

The counterpart of (2.10) for vertical polarization is obtained from (2.10) by replacing E by H and $\Psi(\tau)$ by $\Psi_v(\tau)$, where the subscript v stands for "vertical"; and $\Psi_v(\tau)$ is given by (2.11) when $Ai(u)$ and $Bi(u)$ are replaced by their derivatives with respect to u .

Series for $\Psi(\tau) + 1/\tau$ and $\Psi_v(\tau) + 1/\tau$ which converge for negative (shadow) values of τ are given by Equations (7.31) and (7.53), respectively, with g in place of τ . Table 2.1 gives values of $\Psi(\tau)$ and $\Psi_v(\tau)$ which were obtained from the series for τ negative, and from numerical integration of (2.11) and its analogue for $\tau \geq 0$.

When τ is large and positive Equations (7.35) and (7.55) show that

$$\begin{aligned}\Psi(\tau) + 1/\tau &\sim (i\pi\tau)^{1/2} \exp(-i\tau^3/12), \\ \Psi_v(\tau) + 1/\tau &\sim (i\pi\tau)^{1/2} \exp(i\pi - i\tau^3/12).\end{aligned}\tag{2.12}$$

When τ is large and negative the leading terms in (7.31) and (7.53) give

$$\begin{aligned}\Psi(\tau) + 1/\tau &\sim -i^{1/3} 2.03 \exp[(2.025 + i 1.169)\tau], \\ \Psi_v(\tau) + 1/\tau &\sim -i^{1/3} 3.42 \exp[(0.882 + i 0.509)\tau].\end{aligned}\tag{2.13}$$

Now that we have expressions for the field what do they tell us? For one thing, they may be used to show that the field near the shadow

TABLE 2.1 — VALUES OF $\Psi(\tau)$ AND $\Psi_v(\tau)$

τ	$\Psi(\tau)$	arg. $\Psi(\tau)$	$ \Psi_v(\tau) $	arg. $\Psi_v(\tau)$
3	3.13	-93.5°	3.16	$+104.3^\circ$
2	2.21	-1.8°	2.80	192.4°
1.5	1.945	$+21.6$	2.44	211.3
1.0	1.715	$+32.5$	1.985	219.5
0.5	1.486	34.2	1.522	218.6
0	1.254	30.0	1.089	210.0
-0.5	1.030	22.9	0.724	193.7
-1.0	0.823	15.2	0.459	167.8
-1.5	0.648	8.48	0.317	130.6
-2.0	0.511	$+3.79$	0.281	92.0
-3.0	0.338	$+0.12$	0.288	45.1
-4.0	0.250	-0.12	0.264	22.3
-5.0	0.200	-0.02	0.221	9.71

boundary is almost the same as for a half-plane. Away from the shadow boundary, the field in the shadow can be interpreted as a "crest wave" which reduces to the "edge wave" for a half-plane. The crest wave decreases as an exponential function of ψ in the shadow instead of as $1/\varphi$ in the case of a half-plane. In other words it is much darker behind a parabolic cylinder than behind a half-plane — and the larger the cylinder the darker it is. (A glance at Fig. 2.8 shows that this statement must be qualified for vertical polarization by requiring the observer to be deep in the shadow.) As Figs. 2.7 and 2.8 show, deep in the illuminated region the crest wave behaves like the wave reflected (as computed by geometrical optics) from the illuminated portion of the cylinder.

Now we consider expressions for the surface currents. Let J and J_v be the densities of the conduction currents which flow on the surface of the perfectly conducting parabolic cylinder for the cases of horizontal and vertical polarizations, respectively. J is parallel to E and is perpendicular to the plane of Fig. 2.3 while J_v flows in the plane of the figure. J_v is positive when the current flows in the direction of increasing x . In Section 6 it is shown that when h is large, J and J_v are given approximately by

$$\zeta_0 J \approx \frac{\exp(-ix - i\gamma^3/3)}{\pi h^{1/3}} \int_0^\infty \left[\frac{i^{-2/3} \exp(-i^{1/3} u \gamma)}{Ai(u) - iBi(u)} + \frac{\exp(-iu\gamma)}{Ai(u) + iBi(u)} \right] du, \quad (2.14)$$

$$J_v \approx \frac{i \exp(-ix - i\gamma^3/3)}{\pi} \int_0^\infty \left[-\frac{\exp(-i^{1/3} u \gamma)}{Ai'(u) - iBi'(u)} + \frac{\exp(-iu\gamma)}{Ai'(u) + iBi'(u)} \right] du. \quad (2.15)$$

These expressions are obtained when the relations (13.17) for Airy integrals are used in equations (6.16) and (6.23). Here ζ_0 is the intrinsic impedance of free space. In the rational MKS system which we use $\zeta_0 = 120\pi$ ohms. The factor ζ_0 appears in (2.14) but not in (2.15) because we assume the incident wave for vertical polarization ($H = 1$, $E = \zeta_0 H = 120\pi$) to be 120π times stronger than the one for horizontal polarization ($E = 1$). The primes on $Ai(u)$ and $Bi(u)$ denote their derivatives with respect to u . The parameter γ depends upon the coordinate x of the point at which the current is being observed:

$$\gamma = x/2h^{2/3}. \quad (2.16)$$

Equations (2.14) and (2.15) hold only in the region of the crest of the cylinder and for large h . Under these conditions $x + \gamma^3/3$ is very nearly equal to the distance along the surface measured from the crest of the cylinder.

An expression equivalent to the one for J_v in (2.15) has been derived and tabulated by V. Fock.⁴

Series for $\zeta_0 J$ and J_v which converge for positive (shadow) values of γ are given by equations (6.17) and (6.24). When γ is large and negative the application of the method of steepest descent to integrals (6.16) and (6.23) leads to

$$\begin{aligned}\zeta_0 J &\approx -(x/h)e^{-ix} [1 + i/4\gamma^3 + \cdots], \\ J_v &\approx 2e^{-ix} [1 - i/4\gamma^3 + \cdots],\end{aligned}\tag{2.17}$$

in which $x/h = 2\gamma h^{-1/3}$.

Table 2.2 gives values of $h^{1/3}\zeta_0 J \exp(ix)$ and $J_v \exp(ix)$. The values of J for $\gamma > 0$ were computed from the series, and the ones for $\gamma \leq 0$ were obtained by numerical integration of (2.14). The entries for J_v were taken from the more extensive table given by Fock.⁴ In order to express his results in our terms it is necessary to use the fact that the radius of curvature at the vertex of the parabola is $2h$. A change in the sign of i is also necessary because the time enters Fock's work through $\exp(-i\omega t)$ instead of $\exp(i\omega t)$. The values shown were checked for $\gamma > 0$ by the series and for $\gamma \leq 0$ by numerical integration of (2.15).

Fock's table shows that by the time γ has reached -2 the value of $J_v \exp(ix)$ has become 1.982 at an angle of $+1.45$ degrees. This is close to the limiting value of 2 predicted at $\gamma = -\infty$ by (2.17).

It will be noted that, for large values of h , J is smaller than J_v by

TABLE 2.2 — SURFACE CURRENT DENSITIES

γ	$h^{1/3}\zeta_0 J \exp(ix + i\gamma^3/3)$		$J_v \exp(ix + i\gamma^3/3)$	
	modulus	Argument in degrees	mod.	Arg.
-1.0	2.16	-25.9	1.861	-15.43°
-0.5	1.38	-16.8	1.682	+1.52
0.0	0.77	-30.0	1.399	0.00
0.3	0.515	-44.8	1.197	-6.06
0.6	0.327	-62.9	0.991	-14.23
1.0	0.167	-90.1	0.738	-26.63
1.5	0.066	-125.9	0.488	-42.57
2.0	0.025	-161.6	0.315	-57.98
3.0	0.0033	-230.7	0.130	-87.57
4.0	0.00043	-298.0	0.054	-116.75

the order of $h^{-1/3}$ when γ is of moderate size. It will be shown later that the current density decreases exponentially as one moves into the shadow, and that its rate of decrease is related to that of the field as shown in Fig. 2.6.

We now take up the detailed discussion of the expressions for the field and the current density. It is convenient to consider the current density first. When a plane wave strikes a perfectly conducting plane, the surface current is proportional to the tangential component of H in the incident wave, and flows at right angles to it. In the rational MKS units we are using, the surface current density is two times the incident tangential H . When we consider the illuminated side of a large parabolic cylinder and calculate the current density by doubling the tangential component of the incident H we obtain the approximations

$$\begin{aligned}\zeta_0 J &\approx -2x(4h^2 + x^2)^{-1/2} e^{-ix}, \\ J_v &\approx 2e^{-ix},\end{aligned}\tag{2.18}$$

which hold when x is large and negative. When h is very large but $x/2h$ small these formulas agree with the leading terms of (2.17) which were obtained from our integrals for the current density.

Expressions for the current density deep in the shadow may be obtained from the leading terms of the convergent series by letting γ become large and positive. The exponential decrease is found to be

$$\begin{aligned}|\zeta_0 J| &\approx 1.43h^{-1/3} \exp(-1.013xh^{-2/3}), \\ |J_v| &\approx 1.83 \exp(-0.44xh^{-2/3}),\end{aligned}\tag{2.19}$$

where the numbers appearing in these equations are associated with the smallest zeros of $Ai(u)$ and $Ai'(u)$, respectively. These formulas for a large cylinder are roughly similar to those for propagation over a smooth earth. The radius of curvature at the crest of the cylinder is $2h$. Setting this equal to the radius of the earth gives an exponential rate of decrease for J and J_v which is the same as that over a smooth earth for the two polarizations. Of course, the coefficients multiplying the exponential functions are different. This agreement is not surprising since the Airy integrals are closely related to the Hankel functions of order $1/3$ used in the smooth earth theory.

The surface current densities as a function of the distance $h - y$ below the crest for $h = 1000$ and for $h = 0$ are shown in Fig. 2.4. The equation of the cylinder shows that $h - y = x^2/4h$ so that, from (2.19), $\zeta_0 J$ and J_v decrease in proportion to $h^{-1/3} \exp[-2.025h^{-1/6}(h - y)^{1/2}]$ and $\exp[-.88h^{-1/6}(h - y)^{1/2}]$, respectively, far down in the shadow.

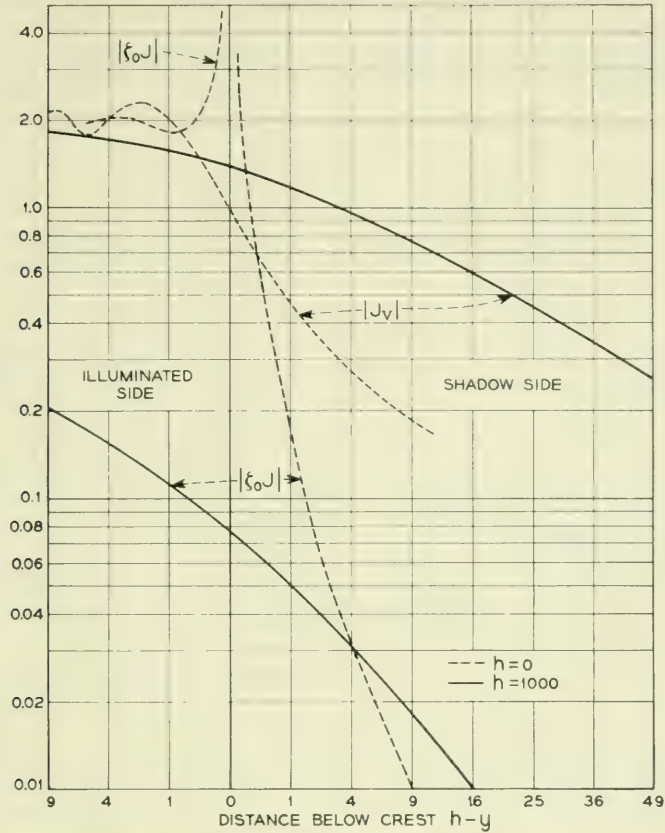


Fig. 2.4 — The surface current density is plotted as a function of the vertical distance below the crest or edge. The curves for $h = 1000$ and $h = 0$ are obtained from Table 2.2 and equations (2.20), respectively. Here, as always, the units are such at $\lambda = 2\pi = 6.28 \dots$

The equations used to compute the curves marked $h = 0$ are

$$\begin{aligned} \xi_0 J &= (i\pi r)^{-1/2} \left[e^{-ir} \mp 2ir^{1/2} \int_{\pm\sqrt{r}}^{\infty} e^{-it^2} dt \right], \\ J_y &= 2(i'\pi)^{1/2} \int_{\pm\sqrt{r}}^{\infty} e^{-it^2} dt, \end{aligned} \tag{2.20}$$

where the upper signs are for the shadow side and the lower ones for the illuminated side. The computations are made easier by the relations

$$\begin{aligned} (\xi_0 J)_- - (\xi_0 J)_+ &= 2, \\ (J_y)_- + (J_y)_+ &= 2, \end{aligned}$$

where the subscripts “-” and “+” denote opposite points on the illuminated side and shadow side, respectively, of the half-plane. These relations follow from (2.20).^{*} The radius vector r from the origin is equal to $-y$ on the half-plane. Equations (2.20) follow quite readily from (2.3). The values of J and J_v for an arbitrary angle of incidence and $h = 0$ are given by expressions (6.6) and (6.22), respectively.

All of the curves in Fig. 2.4, even $|\zeta_0 J|$ for $h = 1,000$, eventually approach the value 2 far down on the illuminated side. It may be shown that $|J_v|$ and $|\zeta_0 J|$ for the half-plane decrease like $(\pi r)^{-1/2}$ and $1/(2\pi^{1/2} r^{3/2})$, respectively, deep in the shadow. Hence as we go to the right in Fig. 2.4, the dashed curves will eventually cross over and lie above the solid curves, which decrease exponentially. The larger h , the lower and flatter is the curve for $|\zeta_0 J|$.

Now we turn to the diffraction field at points far to the right of the cylinder. When $|\psi| \ll 1$, so that we are not too far from the shadow boundary, and h is large, the field is given by (2.10), or by its analogue for vertical polarization. In order to get acquainted with (2.10) we first examine the field when $|\psi| \ll 1$ but $\psi^2 \rho \gg 1$.

When we are so far behind the cylinder that $\psi^2 \rho \gg 1$ even though $|\psi| \ll 1$, the asymptotic expressions (2.4) show that

$$[e^{-ix} + S_1]_\rho \sim \begin{cases} c(\rho)/\psi, & \psi \text{ negative (shadow)} \\ e^{-ix} + \frac{c(\rho)}{\psi}, & \psi \text{ positive} \end{cases} \quad (2.21)$$

Substitution of (2.21) in (2.10) gives

$$E \sim \begin{cases} c(\rho)h^{1/3} \left(\Psi(\tau) + \frac{1}{\tau} \right) \exp(i\tau^3/3), & \psi < 0 \\ c(\rho)h^{1/3} \left(\Psi(\tau) + \frac{1}{\tau} \right) \exp(i\tau^3/3) + e^{-ix}, & \psi > 0 \end{cases} \quad (2.22)$$

The presence of $c(\rho)$ shows that the total wave may be regarded as the sum of the incident wave and a wave spreading out from the crest. The crest wave is the analogue of the edge wave, $c(r)/\varphi$, for the half-plane. In fact, when we are in the region where the $1/\tau$ in (2.22) is the most important term, the crest wave is approximately

$$c(\rho)/\psi \quad (2.23)$$

^{*} They also follow from superposition and consideration of the symmetry of the currents produced on the half-plane $y < 0, x = 0$ when $-A$ is impressed. Here A denotes the system of currents which flows in the upper half-plane $y > 0, x = 0$ when the incident wave falls on a complete plane at $x = 0$.

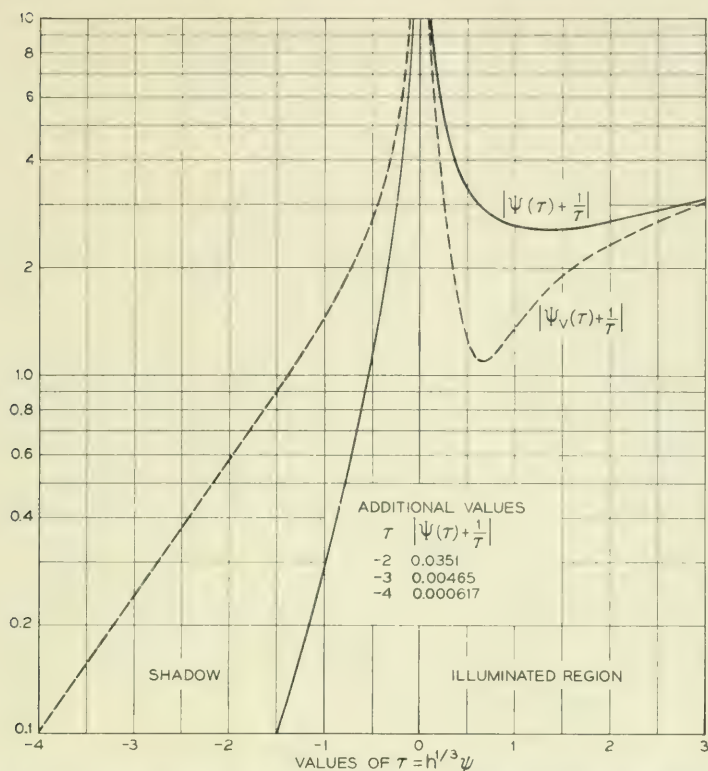


Fig. 2.5 — The amplitude of the crest wave may be obtained from these curves and expressions (2.24). Here $\tau = h^{1/3} \psi$ where ψ is small. However ρ must be large enough to make $\psi^2 \rho \gg 1$.

and this corresponds to a half-plane with its edge at the crest of the cylinder. τ may be small even though we are considering $|\psi|$ to be large enough to make (2.21) and (2.22) hold, i.e. large enough to make $|\psi| \rho^{1/2} \gg 1$. Indeed, multiplying by $h^{1/3}$ gives $|\tau| \rho^{1/2} \gg h^{1/3}$ which may be achieved for small values of τ by making ρ large enough.

It follows from (2.22) and its analogue for vertical polarization that the amplitudes of the crest waves are

$$\begin{aligned}
 (\text{hp}) \quad & (2\pi\rho)^{-1/2} h^{1/3} |\Psi(\tau) + 1/\tau|, \\
 (\text{vp}) \quad & (2\pi\rho)^{-1/2} h^{1/3} |\Psi_v(\tau) + 1/\tau|,
 \end{aligned} \tag{2.24}$$

where the expression (2.9) for $c(\rho)$ has been used. The last factors in (2.24) may be computed from Table 2.1. They are plotted in Fig. 2.5.

When we go deep down into the shadow where τ is large and negative

we see from (2.13) that

$$\begin{aligned} |\Psi(\tau) + 1/\tau| &\sim 2.03 e^{2.025\tau}, \\ |\Psi_v(\tau) + 1/\tau| &\sim 3.42 e^{0.882\tau}, \end{aligned} \quad (2.25)$$

so that the absolute value of the field is

$$\begin{aligned} (\text{hp}) \quad |E| &\sim (2\pi\rho)^{-1/2} h^{1/3} 2.03 \exp(2.025 h^{1/3}\psi), \\ (\text{vp}) \quad |H| &\sim (2\pi\rho)^{-1/2} h^{1/3} 3.42 \exp(0.882 h^{1/3}\psi). \end{aligned} \quad (2.26)$$

where the angle ψ is negative. Thus, as Artmann⁶ has pointed out, the field decreases exponentially as we go into the shadow. The larger h is, the more rapid is the decrease. This supports the statement made earlier that it is darker behind a large cylinder than behind a half-plane.

Comparison of the expressions for the current density and field strength for the shadow regions shows that there is a simple approximate relation between them. Near the crest of the cylinder, where x is small, the radius of curvature is nearly $2h$. Hence the tangent to the parabola drawn from the point P (located at (ρ, ψ) deep in the shadow) touches the parabola at T where x is approximately $-2h\psi$. This is shown in Fig. 2.6. Replacing x by $-2h\psi$ in the expressions (2.19) shows that the current density at T is proportional to the field at P as given by (2.26). It follows that

$$\begin{aligned} (\text{hp}) \quad |E/\zeta_0 J| &\sim 1.41 h^{2/3} (2\pi\rho)^{-1/2} \\ (\text{vp}) \quad |H/J_v| &\sim 1.87 h^{1/3} (2\pi\rho)^{-1/2}. \end{aligned} \quad (2.27)$$

This leads us to picture the field at P as being produced principally by the surface currents around T . The effect of the stronger currents

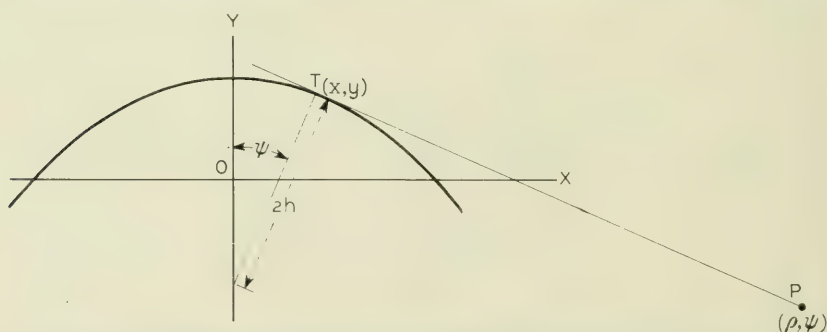


Fig. 2.6 - The field strength at point P (deep in the shadow) specified by the polar coordinates (ρ, ψ) is nearly proportional to the current density at the tangent point T specified by the rectangular coordinates (x, y) .

closer to the crest is perhaps blocked out by the curvature of the cylinder. For comparison with the horizontal polarization case we note that E at (ρ, ψ) for an infinitely long current filament at the origin $\rho = 0$ is given by

$$|E/\zeta_0 I| \sim .5(2\pi\rho)^{-1/2} \quad (2.28)$$

where I is the current carried by the filament and the frequency is such that $\lambda = 2\pi$. There is some difficulty with the picture for vertical polarization because the current element at T points directly towards P and hence should produce very little field there. This is perhaps associated with the fact that the (vp) ratio in (2.27) is smaller than the (hp) ratio by approximately the factor $h^{-1/3}$.

We now leave the shadow region and consider the field at points well inside the illuminated region. Fig. 2.5 shows that for large positive values of τ the amplitude of the crest wave tends to increase with τ . The asymptotic expressions (2.12) show that when τ is large and positive

$$|\Psi(\tau) + 1/\tau| \approx |\Psi_v(\tau) + 1/\tau| \sim (\pi\tau)^{1/2} = (\pi\psi)^{1/2} h^{1/6}, \quad (2.29)$$

and hence the amplitude of the crest wave deep in the illuminated region is, from (2.22) and its analogue,

$$\begin{aligned} \text{(hp)} \quad & |E - e^{-ix}| \sim (2\pi\rho)^{-1/2} (\pi\psi h)^{1/2}, \\ \text{(vp)} \quad & |H - e^{-ix}| \sim (2\pi\rho)^{-1/2} (\pi\psi h)^{1/2}. \end{aligned} \quad (2.30)$$

Since (2.30) is derived from the general expression (2.10) it is subject to the restrictions mentioned just below equation (2.11). In particular the angle ψ should be small (but we must still have $\psi^2\rho \gg 1$ as assumed in (2.22)). When ψ is positive, an application of the laws of geometrical optics to determine the reflection from the curved surface of the parabolic cylinder leads to the expressions¹²

$$\begin{aligned} \text{(hp)} \quad & |E - e^{-ix}| \sim \left[\frac{h \tan(\psi/2)}{\rho} \right]^{1/2} \sec(\psi/2), \quad \psi > 0 \\ \text{(vp)} \quad & |H - e^{-ix}| \sim \left[\frac{h \tan(\psi/2)}{\rho} \right]^{1/2} \sec(\psi/2), \quad \psi > 0 \end{aligned} \quad (2.31)$$

for the reflected wave. When ψ is small these expressions reduce to (2.30) as they should.

Expressions (2.31) may also be obtained from our analysis by start-

¹² In our two-dimensional case the calculation of the required radius of curvature, etc., is not difficult. General theorems dealing with problems of this sort and references to earlier work are given in the paper, A General Divergence Formula, H. J. Riblet and C. B. Barker, J. Appl. Phys. **19**, pp. 63-70, 1948.

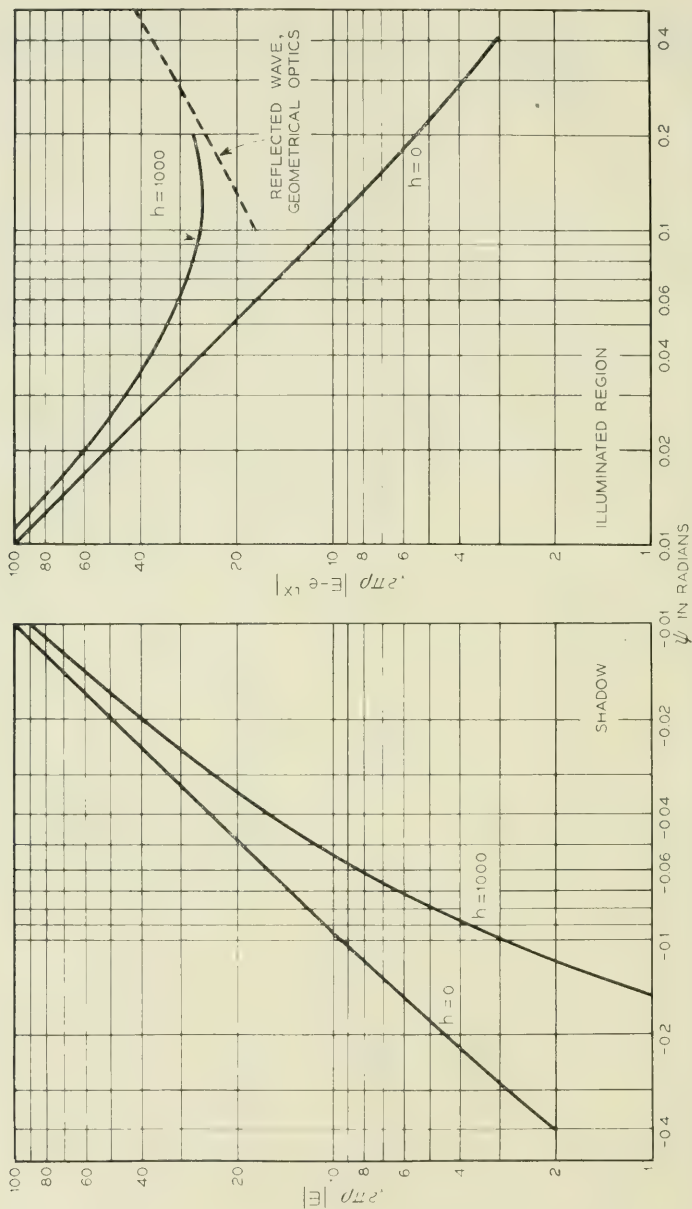


Fig. 2.7 — Strength of crest wave for horizontal polarization. The curves fail for $|\psi| < \rho^{-1/2}$, and the $h = 1000$ curve becomes less reliable as ψ increases. The exponential decrease deep in the shadow is given by equation (2.26). The dashed curve for the reflected wave is computed from equation (2.31).

ing with equations (7.36) and (7.56). Furthermore, it may be verified that the phase angles of the reflected waves as computed from (7.36) and (7.56) agree with those computed from geometrical optics when reflection coefficients of -1 and $+1$ are assumed for hp and vp, respectively.

The amplitudes of the crest waves for $h = 1,000$ and $h = 0$ are shown for hp in Fig. 2.7 and for vp in Fig. 2.8. Of course, when $h = 0$ the crest wave reduces to the edge wave from a half-plane. The curves for $h = 1,000$ were computed from equations (2.24) and the curves of Fig. 2.5 (or their equivalent when τ is small). The curves for $h = 0$ were computed from

$$|E| \sim (2\pi\rho)^{-1/2} \left| \frac{1}{2 \sin(\psi/2)} + \frac{1}{2 \cos(\psi/2)} \right|, \quad \psi < 0$$

(hp)

$$|E - e^{-ix}| \sim (2\pi\rho)^{-1/2} \left| \frac{1}{2 \sin(\psi/2)} + \frac{1}{2 \cos(\psi/2)} \right|, \quad \psi > 0 \quad (2.32)$$

$$|H| \sim (2\pi\rho)^{-1/2} \left| \frac{1}{2 \sin(\psi/2)} - \frac{1}{2 \cos(\psi/2)} \right|, \quad \psi < 0$$

(vp)

$$|H - e^{-ix}| \sim (2\pi\rho)^{-1/2} \left| \frac{1}{2 \sin(\psi/2)} - \frac{1}{2 \cos(\psi/2)} \right|, \quad \psi > 0$$

which follow from (2.1), (2.2), (2.4) and (2.6).

From equation (2.21) onward we have been discussing the field for values of ψ and ρ such that $\rho\psi^2 \gg 1$. For these values the concept of the crest wave is helpful in visualizing the behavior of the field. Now we consider the field at points close to the boundary of the geometric shadow far behind the cylinder. This is the region in which Artmann⁶ was especially interested. His results for the shift of the field may be obtained from (2.10) and its analogue by taking $|\psi|$ to be very small.

At the shadow boundary $\psi = 0$ and $[\exp(-ix) + S_1]_\rho = 1/2$. Hence the region of interest at present is in the neighborhood of the point $t_1 = 0$, $|\exp(-ix) + S_1| = 1/2$ of Fig. 2.2. A magnified view of this region showing the shift of the field is given in Fig. 2.9. The figure shows that, for a given value of $\rho\psi$, $|E|$ for hp is less than $|H|$ for vp. As Artmann has pointed out, this is to be expected since the reflection coefficient for E (hp) is roughly -1 and the reflected wave therefore

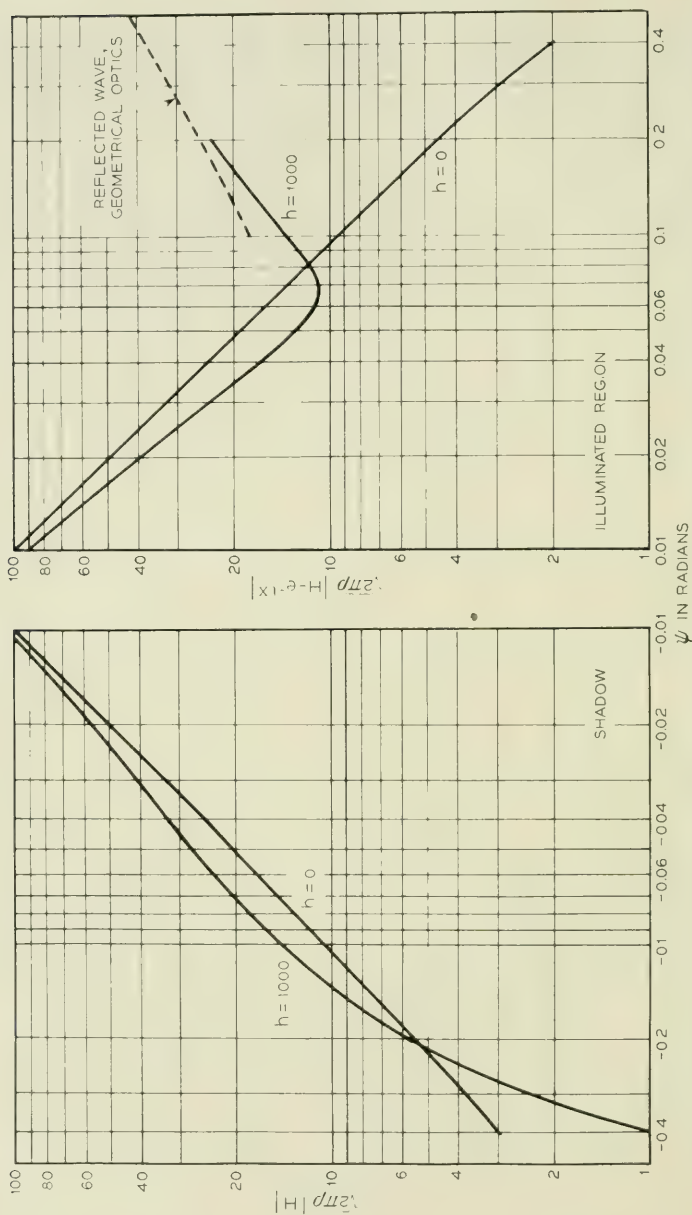


Fig. 2.8 — Strength of crest wave for vertical polarization. The remarks made in the caption of Fig. 2.7 also hold here. Incidentally, a value of $h = 1000$ corresponds to a radius of curvature of $2000/2\pi = 318$ wavelengths at the crest.

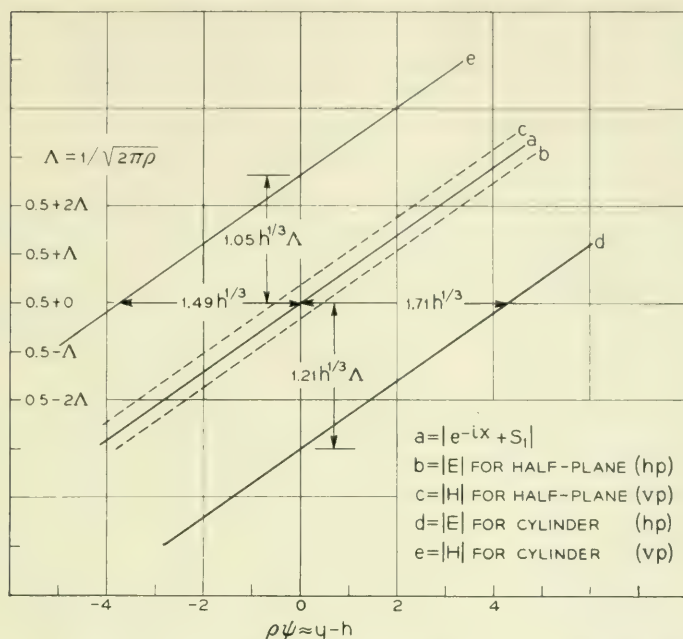


Fig. 2.9 - Behavior of $|E|$ and $|H|$ on the shadow boundary far behind the half-plane or parabolic cylinder. This is a magnified view of the region around $t_1 = 0$, $|\exp(-ix) + S_1| = \frac{1}{2}$ in Fig. 2.2.

tends to cancel the direct wave when ψ is very small. On the other hand, the reflection coefficient for H (vp) is $+1$ and the reflected wave tends to add to the direct wave.

The distances $1.71h^{1/3}$ and $-1.49h^{1/3}$ appearing in Fig. 2.9 are the amounts, measured in units for which the wavelength is 2π , by which $|E|$ and $|H|$ are shifted by the curvature of the parabolic cylinder. If $y' - h'$ is the shift in meters for $|E|$ and if the radius of curvature of the crest is $a = 2h'$ meters, Fig. 2.9 gives $\beta(y' - h') = 1.71(\beta h')^{1/3} = 1.71(\beta a/2)^{1/3}$ where $\beta = 2\pi/\lambda$. Thus $y' - h' = 0.399\lambda(a'/\lambda)^{1/3}$ meters. The corresponding shift for $|H|$ is $-0.346\lambda(a'/\lambda)^{1/3}$ meters. Artmann gives the values 0.39 and -0.20 for the respective coefficients. The discrepancy between -0.346 and -0.20 is cause for worry because it seems to indicate either a mistake in our work, which I have been unable to locate, or a shortcoming in the approximations made by Artmann for the case of vertical polarization.

As h approaches zero the parabolic cylinder becomes a half-plane and the curves d and e should approach curves b and c , respectively. According to Fig. 2.9 both d and e approach curve a . This failure is an

indication of the errors introduced by the approximations used in the derivation of (2.10) and its analogue.

3. RADIO PROPOGATION OVER A SUCCESSION OF RIDGES ON THE EARTH'S SURFACE

The results mentioned in Section 1 concerning propagation over a succession of ridges may be obtained from the expressions and curves of Section 2 as follows: Consider the situation shown in Fig. 3.1. Let a radio wave start out from a transmitter at T . We assume that by the time it arrives at the first ridge at P it has become equivalent to a plane wave of amplitude A/ℓ traveling in the direction TP , where A is a constant depending on the strength of the transmitter. For the sake of simplicity the waves reflected from the ground are neglected. In a more careful study they would have to be included.*

In order to calculate the strength of the wave at the second ridge, we assume it to be a crest wave coming from P . Let G denote the value of $|E|$ (we assume the case of horizontal polarization since the reflection coefficient of physical materials approaches -1 for almost grazing incidence) at Q corresponding to a plane wave of unit amplitude incident on P . From Fig. 3.1 we see that the values of ψ and ρ to be used in computing G are $\psi \approx -\ell/R$, $\rho = 2\pi\ell/\lambda$, λ = wavelength, R = radius of earth. The value of h depends upon the radius of curvature of the ridge: $2h = 2\pi$ (radius of curvature)/ λ .

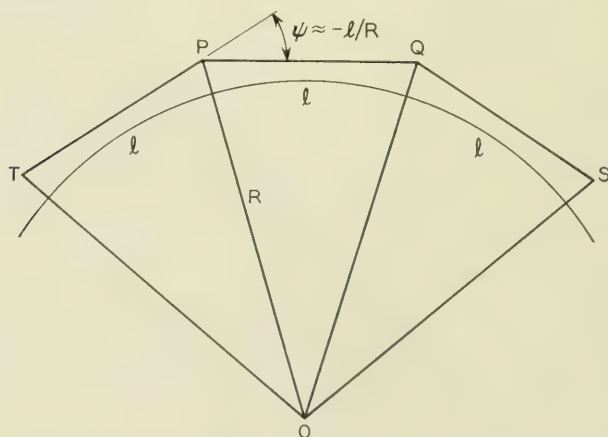


Fig. 3.1 — Diagram showing ridges at P and Q which diffract the radio wave starting from T so that a portion of it is received at S .

* A method for doing this (for one hill) is given in Reference 1, page 417.

The amplitude of the wave striking the ridge at Q is $AG/\ell\sqrt{2}$. The $\sqrt{2}$ comes from the horizontal sidewise spreading of the wave in going from ℓ to 2ℓ . If we were dealing with the energy instead of the amplitude, the factor would be 2 instead of $\sqrt{2}$. When this wave is assumed to be plane and traveling in the PQ direction, similar reasoning shows that the amplitude of the disturbance at the receiver S is $AG^2/\ell\sqrt{3}$.

If, instead of two ridges at P and Q , as shown in Fig. 3.1, there are N ridges between the transmitter at T and the receiver at S , the amplitude of the radio wave at S is $AG^N/\ell\sqrt{N+1}$. The distance between T and S is approximately $(N+1)\ell$, and the free space amplitude at S is $A/(N+1)\ell$. Hence

$$\frac{\text{Actual Amplitude at } S}{\text{Free Space Amplitude at } S} = G^N(N+1)^{1/2}. \quad (3.1)$$

The actual field at S is therefore

$$20 N \log_{10} (1/G) - 10 \log_{10} (N+1) \quad (3.2)$$

db below the free space field.

As an example, let us assume a distance of 280 miles between the transmitter and receiver, and a distance of 40 miles between successive ridges. This gives $N = 6$. For a wavelength of 10 meters and a radius of curvature of 100 meters for the diffracting ridges, the formulas of Section 2 show that the ridges behave like half-planes and that $G \approx 0.227$. Equation (3.2) then says that, for a distance of 280 miles and a wavelength of 10 meters the actual field should be about 69 db below the free space field. Although this is in fair agreement with the experimental results, calculations for other distances indicate that the field strengths predicted by (3.2) tend to be smaller than the ones observed.

When the work is carried through for $\lambda = 1$ meter and a distance of 280 miles, (3.2) says that the field is 120 db below free space. The observed fields are 70 ± 15 db below the free space value.

These figures suggest that the roughness of the earth's surface might possibly account for transmission far beyond the horizon for wavelengths of the order of 10 meters. For wavelengths of the order of 1 meter either the approximations leading to (3.2) break down or some other explanation is required.

4. SERIES FOR THE ELECTROMAGNETIC FIELD

Here we set down series for the electromagnetic field when a plane wave strikes a perfectly conducting parabolic cylinder. Since Epstein's

classical work deals with the general case of finite conductivity, the series we use are special cases of the ones discussed by him.

The parabolic coordinates (ξ, η) which we shall use are related to the rectangular coordinates (x, y) and polar coordinates (r, φ) as follows:

$$\begin{aligned} x + iy &= (\xi + i\eta)^2/2i = r e^{i\varphi}, \\ x &= \xi\eta = r \cos \varphi, & y &= (\eta^2 - \xi^2)/2 = r \sin \varphi, \\ r &= (x^2 + y^2)^{1/2} = (\xi^2 + \eta^2)/2, \\ dx^2 + dy^2 &= (\xi^2 + \eta^2) (d\xi^2 + d\eta^2) = 2r(d\xi^2 + d\eta^2), \\ \xi &= (2r)^{1/2} \cos(\varphi/2 + \pi/4), \\ \eta &= (2r)^{1/2} \sin(\varphi/2 + \pi/4). \end{aligned} \tag{4.1}$$

The lines $\eta = \text{constant}$ are a series of downward-curving confocal parabolas having their focus at the origin. The parabolic cylinder $x^2 = 4h(h - y)$ is given by $\eta^2 = 2h$. This special value of η will be called η_0 :

$$\eta_0 = (2h)^{1/2} \geq 0, \quad h = \eta_0^2/2. \tag{4.2}$$

When $\eta_0 = 0$, the cylinder reduces to the half-plane $x = 0, y \leq 0$. The lines $\xi = \text{constant}$ are halves of upward-curving confocal parabolas having their common focus at the origin. Outside the cylinder $\eta > \eta_0 \geq 0$, so η is always positive in our work. ξ is positive in the half-plane $x > 0$ and negative in $x < 0$.

For much of our work we shall assume the incident wave to come in at the angle $\theta, 0 \leq \theta \leq \pi$, as shown in Fig. 4.1. As mentioned in Section

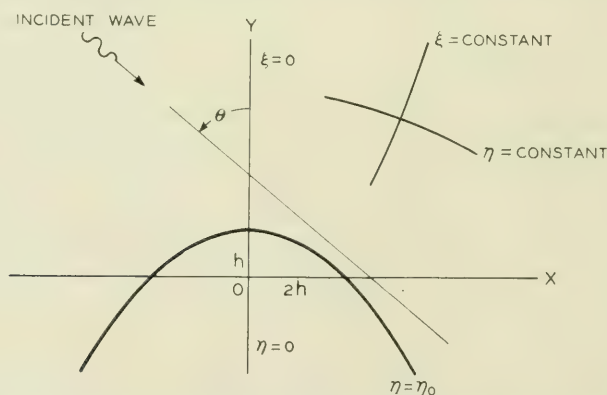


Fig. 4.1 — This diagram shows the angle θ of the incident wave and the surface of the perfectly conducting cylinder $x^2 = 4h(h - y)$ (or $\eta = \eta_0$).

2, the field quantities are assumed to depend upon the time through the factor $\exp(i\omega t)$ where ω is the radian frequency.

The wave equations for horizontal and vertical polarization are, respectively,

$$\frac{\partial^2 E}{\partial \xi^2} + \frac{\partial^2 E}{\partial \eta^2} + (\xi^2 + \eta^2)E = 0 \quad (4.3)$$

$$\frac{\partial^2 H}{\partial \xi^2} + \frac{\partial^2 H}{\partial \eta^2} + (\xi^2 + \eta^2)H = 0 \quad (4.4)$$

where, as explained in Section 2, the unit of length has been chosen so that the wavelength $\lambda = 2\pi$. On the surface of the perfectly conducting cylinder, i.e. for $\eta = \eta_0$, we must have $E = 0$ and $\partial H / \partial \eta = 0$. When E and H are known the remaining components of the field may be computed from Maxwell's equations.

Special solutions of (4.3) (and (4.4)) are

$$\exp[i(\eta^2 - \xi^2)/2] U_n(\xi i^{1/2}) U_n(\eta i^{-1/2}), \quad (4.5)$$

$$\exp[i(\eta^2 - \xi^2)/2] U_n(\xi i^{1/2}) W_n(\eta i^{-1/2}), \quad (4.6)$$

where $i^{1/2}$ stands for $\exp(i\pi/4)$ and $U_n(z)$, $W_n(z)$ satisfy the equation

$$\frac{d^2 T_n(z)}{dz^2} - 2z \frac{dT_n(z)}{dz} + 2nT_n(z) = 0. \quad (4.7)$$

Another solution of (4.7) with which we shall be concerned is $V_n(z)$. These three solutions are defined by contour integrals of the form

$$(2\pi i)^{-1} \int \exp[f(t)] dt \quad \text{where} \quad f(t) = -t^2 + 2zt - (n+1) \log t.$$

The path of integration for $U_n(z)$ comes in from $-\infty$ where $\arg t = -\pi$, encircles the origin counterclockwise and runs out to $-\infty$ with $\arg t = \pi$. The path for $V_n(z)$ runs from $-\infty$ where $\arg t = \pi$ to $+\infty$ where $\arg t = 0$, and the path for $W_n(z)$ runs from $+\infty$ to $-\infty$ where $\arg t = -\pi$. The integrals are written at greater length in equations (9.1) and the paths of integration are shown in Fig. 9.1. Since the paths may be joined to form a closed path containing no singularities of the integrand it follows that

$$U_n(z) + V_n(z) + W_n(z) = 0. \quad (4.8)$$

When n is a non-negative integer

$$U_n(z) = H_n(z)/n! = \frac{(-)^n}{n!} e^{z^2} \frac{d^n}{dz^n} e^{-z^2} \quad (4.9)$$

where $H_n(z)$ is Hermite's polynomial. When z becomes very large the leading terms in (9.17) and (9.16) give

$$U_n(z) \sim (2z)^n/n!, \quad (4.10)$$

$$W_n(\eta i^{-1/2}) \sim i(i^{1/2}/\eta)^{n+1} e^{-i\eta^2/2} \pi^{1/2}. \quad (4.11)$$

In order to obtain a series for the incident wave

$$\exp[-ix \sin \theta + iy \cos \theta]$$

shown in Fig. 4.1 we consider the special case $\theta = 0$. In this case the wave is simply $\exp(iy)$ or $\exp[i(\eta^2 - \xi^2)/2]$ and may be obtained by setting $n = 0$ in (4.5). This suggests that the incident wave may be expressed as the sum of terms like (4.5). The series turns out to be

$$\begin{aligned} & \exp[-ix \sin \theta + iy \cos \theta] \\ &= \exp[-i\xi\eta \sin \theta + i \cos \theta(\eta^2 - \xi^2)/2] \\ &= \exp[-izz' \sin \theta - \cos \theta(z^2 + z'^2)/2] \quad (4.12) \\ &= e^{iy \sec(\theta/2)} \sum_{n=0}^{\infty} n! (-iw/2)^n U_n(z) U_n(z') \end{aligned}$$

where

$$\begin{aligned} w &= \tan(\theta/2), \\ z &= \xi i^{1/2}, \\ z' &= \eta i^{-1/2}. \end{aligned} \quad (4.13)$$

This series has been studied by a number of writers. It goes back to Mehler¹³ who obtained it by evaluating the integral

$$\pi^{-1/2} e^{x^2} \int_{-\infty}^{\infty} \exp[-(t - iy)^2 - (x + iat)^2] dt$$

first in closed form, and then as a series (by using the generating function $\exp[-(-iat)^2 + 2(-iat)x]$ for $H_n(x)$ and integrating termwise). This leads to

$$\begin{aligned} (1 - a^2)^{-1/2} \exp \left[\frac{2xya - (x^2 + y^2)a^2}{1 - a^2} \right] \\ = \sum_0^{\infty} H_n(x) H_n(y) a^n / 2^n n! \end{aligned} \quad (4.14)$$

which is equivalent to (4.12). Since (4.14) converges when $|a| < 1$, (4.12) converges when $|w| < 1$ or $|\theta| < \pi/2$.

¹³ Reihenentwicklungen nach Laplaceschen Functionen höher Ordnung, J. Reine Angew., Math., **66**, pp. 161-176, 1866.

When the incident wave strikes the cylinder the reflected wave has some of the characteristics of a wave spreading radially outwards. Such a wave contains the factor $\exp(-ir) = \exp[-i(\xi^2 + \eta^2)/2]$. Consideration of the exponential factors in (4.6) and (4.11) suggests that the reflected wave may be expressed as the sum of terms of the form (4.6). The coefficients in this series are to be determined so that $E = 0$ or $\partial H/\partial \eta = 0$ at the surface $\eta = \eta_0$, the incident wave being represented by (4.12).

For the case of horizontal polarization this procedure gives

$$E = e^{iy} \sec(\theta/2) \sum_0^\infty n!(-iw/2)^n U_n(z) [U_n(z') - W_n(z')U_n(z'_0)/W_n(z'_0)] \quad (4.15)$$

$$E = \exp[-ix \sin \theta + iy \cos \theta] - e^{iy} \sec(\theta/2) \sum_0^\infty n!(-iw/2)^n U_n(z) W_n(z') U_n(z'_0)/W_n(z'_0) \quad (4.16)$$

for the complete field. These are special cases of Epstein's results. Here z'_0 is the value of z' which corresponds to the surface of the cylinder:

$$z'_0 = i^{-1/2} \eta_0 = (2h/i)^{1/2} \quad (4.17)$$

The entries for regions II and II' (these are regions in the m -plane ($m = n + 1$) which, as Figs. 11.2 and 12.2 show, contain the large positive values of n) in Tables 12.2 and 12.4 of Section 12 may be used to show that as $n \rightarrow \infty$

$$\begin{aligned} V_n(z'_0)/W_n(z'_0) &\sim i^{-2n} \exp[2\eta_0(2in)^{1/2}], \\ U_n(z'_0)/W_n(z'_0) &= -1 - V_n(z'_0)/W_n(z'_0), \\ U_n(z)W_n(z') &\sim -\frac{\exp[-iy - n(2n/i)^{1/2}]}{4[\Gamma(1 + n/2)]^2} \\ &\quad \{ \exp[-\xi(2n/i)^{1/2}] + i^{2n} \exp[\xi(2n/i)^{1/2}] \}. \end{aligned} \quad (4.18)$$

Since

$$n!/[\Gamma(1 + n/2)]^2 \sim 2^n (\pi n/2)^{-1/2}$$

the series in (4.15) and (4.16) converge if $|w| = |\tan \theta/2| < 1$.

Series for H similar to those of (4.15) and (4.16) may be obtained for the case of vertical polarization. The boundary condition at $\eta = \eta_0$ is now $\partial H/\partial \eta = 0$. It is convenient to introduce the functions $'U_n(z)$, $'V_n(z)$, $'W_n(z)$ defined by

$$'U_n(z) = -zU_n(z) + \partial U_n(z)/\partial z \quad (4.19)$$

and the two other equations obtained when U is replaced by V and W . The prime is placed in front of U instead of behind to avoid mistaking $'U_n(z)$ for $\partial U_n(z)/\partial z$. The function $'U_n(z')$ makes its appearance when $\partial H/\partial \eta$ is calculated for the boundary condition. Since $iy = -(z^2 + z'^2)/2$ we have

$$\frac{\partial}{\partial \eta} e^{iy} U_n(z') = i^{-1/2} e^{iy} 'U_n(z'). \quad (4.20)$$

The analogues of (4.15) and (4.16) for vertical polarization are (assuming now that H for the incident wave is of unit amplitude).

$$H = e^{iy} \sec(\theta/2) \sum_0^\infty n! (-iw/2)^n U_n(z) [U_n(z') - W_n(z')'U_n(z'_0) / 'W_n(z'_0)], \quad (4.21)$$

$$= \exp[-ix \sin \theta + iy \cos \theta] - e^{iy} \sec(\theta/2) \sum_0^\infty n! (-iw/2)^n U_n(z) W_n(z')'U_n(z'_0) / 'W_n(z'_0), \quad (4.22)$$

and these series converge if $|w| < 1$.

If the parabolic cylinder is merely a good conductor, instead of being perfect, the boundary conditions at $\eta = \eta_0$ are approximately¹⁴ $E = -\zeta H_\xi$, $E_\xi = \zeta H$. Here E_ξ and H_ξ denote the ξ components of the electric and magnetic intensities and ζ is the intrinsic impedance

$$\zeta = [i\omega\mu/(g + i\omega\epsilon)]^{1/2} \quad (4.23)$$

of the cylinder material. ζ is assumed to be small compared to the intrinsic impedance $\zeta_0 = (\mu_0/\epsilon_0)^{1/2} = \omega\mu_0$ (since $\lambda = 2\pi$) of the external medium. In these expressions μ , ϵ , g are the permeability, dielectric constant, and conductivity of the cylinder; and μ_0 and ϵ_0 refer to the external medium.

When we set

$$\begin{aligned} \sigma &= i^{-1/2}(\xi^2 + \eta_0^2)^{1/2} \zeta_0/\zeta, \\ \tau &= i^{-1/2}(\xi^2 + \eta_0^2)^{1/2} \zeta/\zeta_0, \end{aligned} \quad (4.24)$$

the boundary condition for hp becomes $\sigma E = -\partial E/\partial z'$ at $z' = z'_0$. When σ is assumed to be constant we obtain

$$E = e^{iy} \sec(\theta/2) \sum_0^\infty n! (-iw/2)^n U_n(z) \{U_n(z') - W_n(z')[\sigma U_n(z'_0) + 'U_n(z'_0)] / [\sigma W_n(z'_0) + 'W_n(z'_0)]\} \quad (4.25)$$

¹⁴ *Electromagnetic Waves*, S. A. Schelkunoff, D. Von Nostrand Co., N. Y. (1943) p. 89. See also G. A. Hufford, *Quart. Appl. Math.* **9**, pp. 391-403, 1952, where reference is made to the work of Leontovich and Fock.

which reduces to (4.15) when $\sigma \rightarrow \infty$. The constancy of σ may be achieved by either taking the properties of the cylinder material to change in a suitable way or, roughly, by taking η_0^2 (and hence h) to be so large that only the nearly constant values of σ at the crest of the cylinder have an effect on the result (assuming $\theta = \pi/2$, and restricting our attention to the region near the shadow boundary).

The corresponding expression for vertical polarization may be obtained from (4.25) by replacing E and σ by H and τ , respectively. We shall refer to (4.25) and its analogue later in connection with the field in the shadow (Section 7) and with Fock's⁵ investigation of the surface currents on gently curved conductors (Section 6).

5. INTEGRALS FOR THE FIELD

When the curvature of the cylinder is small, i.e., when h is many wavelengths, the series of Section 4 converge slowly. The work initiated by G. N. Watson¹⁵ on the smooth earth problem suggests that we convert the series into contour integrals with n as the complex variable of integration. When this is done we get an integral with the path of integration L_1 shown in Fig. 5.1. Thus, for example, expression (4.16) for E is transformed into

$$E = \exp[-ix \sin \theta + iy \cos \theta] - \frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_1} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z') U_n(z'_0) dn / W_n(z'_0). \quad (5.1)$$

At first sight it seems that not much can be done with this integral because the integral obtained by deforming L_1 into L_2 does not converge (this is explained in the discussion of Table 5.1). However, some ex-

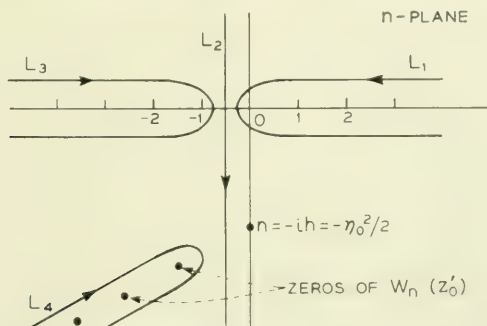


Fig. 5.1 — Paths of integration in the complex n -plane.

¹⁵ Proc. Roy. Soc., London (A) **95**, p. 83, 1918.

perimentation shows that if we set $U_n(z'_0) = -V_n(z'_0) - W_n(z'_0)$ in (4.16), the series splits into two series, one of which may be summed and the other may be converted into an integral along the path L_2 of Fig. 5.1:

$$E = \exp[-ix \sin \theta + iy \cos \theta] + S_1 + S_2(h),$$

$$S_1 = e^{iy \sec(\theta/2)} \sum_0^{\infty} n! (-iw/2)^n U_n(z) W_n(z'), \quad (5.2)$$

$$S_2(h) = e^{iy \sec(\theta/2)} \sum_0^{\infty} n! (-iw/2)^n U_n(z) W_n(z') V_n(z'_0) / W_n(z'_0).$$

The series for S_1 may be summed by replacing $U_n(z)$ and $W_n(z')$ by their expressions (9.19) in terms of definite integrals, and interchanging the order of summation and integration. The resulting series may be summed and the integrations performed. The result is Sommerfeld's integral for a diffracted plane wave:

$$S_1 = -(i/\pi)^{1/2} e^{ix \sin \theta + iy \cos \theta} \int_{T_1}^{\infty} e^{-it^2} dt, \quad (5.3)$$

$$T_1 = \eta \cos \frac{\theta}{2} - \xi \sin \frac{\theta}{2} = (2r)^{1/2} \sin \left(\frac{\varphi - \theta}{2} + \frac{\pi}{4} \right).$$

The inversion of the order of summation and integration may be justified when $|w| = |\tan(\theta/2)| < 1$ (in which case the series in (5.2) converge) by using (1) the result that $|R_N| < |a^N/N!|$ when a is real in

$$\sum_0^{N-1} \frac{(ia)^n}{n!} = e^{ia} - R_N,$$

and (2) the inequality

$$\left[\int_0^{\infty} t^N \exp[-t^2 + 2^{1/2}bt] dt \right]^2$$

$$< \int_0^{\infty} e^{-2(1-\alpha)t^2} t^{2N} dt \int_{-\infty}^{\infty} \exp[-2\alpha t^2 + 2^{3/2}bt] dt$$

$$< A 2^{-N} b^{-1/2} N^{-1/4} N! \exp(2bN^{1/2})$$

This inequality holds when $N \gg b^2$ and at the same time $N \gg 1$. The value of A is independent of N and b is a number which exceeds $|\xi|$. In this work the parameter α has been arbitrarily introduced; and has then been chosen so as to make the product of the two integrals a minimum when N is large. This value of α is $bN^{-1/2}$.

When the series (5.2) for $S_2(h)$ is converted into a contour integral taken along the path L_1 , by the procedure used to obtain (5.1), it is

seen that L_1 may be deformed into L_2 and we obtain

$$S_2(h) = \frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2} \right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z') U_n(z'_0) dn W_n(z'_0) \quad (5.4).$$

Whether a particular integrand, such as the one shown in (5.4), converges at the ends $n = \pm i\infty$ of L_2 can often be decided from Table 5.1. This table gives a rough idea of the behavior of the various functions in terms of powers of i . For example, if the integrand should turn out to be proportional to $i^n = \exp(i\pi n/2)$ at $n = i\infty$, the integral will converge like $\exp(-\pi |n|/2)$.

TABLE 5.1

Function	Order of Magnitude — Rough Approximation	
	near $n = i\infty$	near $n = -i\infty$
i^n	0	∞
i^{-n}	∞	0
$\sin \pi n$	i^{-2n}	i^{2n}
$\Gamma(n+1)$	i^n	i^{-n}
$U_n(z)$	$i^{-3n/2}$	$i^{3n/2}$
$V_n(z)$	$i^{-3n/2}$	$i^{-n/2}$
$W_n(z)$	$i^{n/2}$	$i^{3n/2}$

The approximations for $\Gamma(n+1)$ follow from its asymptotic expression, and those for the parabolic cylinder functions come from Tables 12.2 and 12.4. The entries for the cylinder functions may also be surmised from expressions (9.4) which hold for $z = 0$.

Table 5.1 may be used to show that the integrand in (5.1) is of the order of i^n as $n \rightarrow -i\infty$. Hence there is no hope of deforming L_1 into L_2 in this case. On the other hand, the integrand in (5.4) is of the order of i^n as $n \rightarrow i\infty$ and of i^{-n} as $n \rightarrow -i\infty$, and therefore (5.4) converges exponentially. In fact, it converges for all real positive values of $w = \tan \theta/2$. This enables us to obtain an expression for the field which holds for $0 < \theta < \pi$ (i.e. it is not subject to the restriction $|w| < 1$ required by (4.16)). This expression, which is fundamental for our work, has the form

$$E = \exp[-ix \sin \theta + iy \cos \theta] + S_1 + S_2(h). \quad (5.5)$$

Here S_1 and $S_2(h)$ are given by (5.3) and (5.4), respectively.

In working with (5.5) it is sometimes convenient to use the expression

$$\exp [-ix \sin \theta + iy \cos \theta] + S_1$$

$$= (i/\pi)^{1/2} \exp [-ix \sin \theta + iy \cos \theta] \int_{-\infty}^{T_1} e^{-it^2} dt \quad (5.6)$$

which follows from (5.3) and

$$\int_{-\infty}^{\infty} \exp (-it^2) dt = (\pi/i)^{1/2}. \quad (5.7)$$

The development leading to (5.5) shows that it satisfies the boundary condition $E = 0$ at $\eta = \eta_0$ for $0 < w < 1$. That (5.5) also satisfies the condition for the extended range $0 < w < \infty$ follows immediately from

$$\begin{aligned} & \frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2} \right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) V_n(z') dn \\ &= -(i/\pi)^{1/2} \exp [-ix \sin \theta + iy \cos \theta] \int_{-\infty}^{T_1} \exp (-it^2) dt \\ &= -\exp [-ix \sin \theta + iy \cos \theta] - S_1 \end{aligned} \quad (5.8)$$

when we note that setting $z' = z'_0$ reduces $S_2(h)$ to the left hand side of (5.8) (with $z' = z'_0$).

Equation (5.8) is due to T. M. Cherry¹⁶ who obtained it by expressing the cylinder functions as integrals and interchanging the order of integration (he works with the function $D_n(z)$ of our equations (9.2)). Substituting the integrals (9.19) for $U_n(z)$ and $V_n(z')$ in (5.8) and interchanging the order of integration leads to a similar derivation. Equation (5.8) may also be obtained by deforming L_2 into L_1 when $0 < w < 1$ and into L_3 when $1 < w < \infty$. This leads to the two series

$$e^{iy} \sec \frac{\theta}{2} \sum_{n=0}^{\infty} (-iw/2)^n n! U_n(z) V_n(z'), \quad (5.9)$$

$$-e^{iy} \sec \frac{\theta}{2} \sum_{n=-1}^{-\infty} (-iw/2)^n [\Gamma(n+1) U_n(z)] V_n(z') \quad (5.10)$$

which may be summed in much the same way as was (5.2) for S_1 .

An expression for E which is useful in the study of the current density on the surface of the cylinder may be obtained from (5.5) by combining expression (5.8) for $\exp [-ix \sin \theta + iy \cos \theta] + S_1$ with expression (5.4)

¹⁶ Expansions in Terms of Parabolic Cylinder Functions, Proc. Edinburgh Math. Soc., Ser. 2, **8**, pp. 50-65, 1948.

for $S_2(h)$:

$$E = \frac{e^{i\eta} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z') \left[\frac{V_n(z'_0)}{W_n(z'_0)} - \frac{V_n(z')}{W_n(z')} \right] dn. \quad (5.11)$$

When w exceeds unity (or when $w = 1$ and $\xi > \eta \geq \eta_0 \geq 0$) in (5.11), it may be verified with the help of Tables 12.2 and 12.4 that L_2 may be deformed into $L_3 + L_4$. When n is a negative integer the quantity within the brackets in (5.11) vanishes because of (4.8) and because $U_n(z) = 0$. The contribution of L_3 is zero since it encloses no poles. The contribution of L_4 is equal to the sum of the residues at the poles given by $W_n(z'_0) = 0$. Hence, when $w > 1$,

$$E = -\pi e^{i\eta} \sec \frac{\theta}{2} \sum_{s=1}^{\infty} \left[\left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1) U_n(z) W_n(z') V_n(z'_0)}{\sin \pi n \partial W_n(z'_0) / \partial n} \right]_{n=n_s} \quad (5.12)$$

where $n = n_s$ is the s th zero of $W_n(z'_0)$. This series also converges when $w = 1$ and $\xi > \eta \geq \eta_0 \geq 0$ (which is roughly the shadow region). The preceding inequality does not necessarily specify the complete region of convergence.

Cherry¹⁶ has also pointed out that the expression for a plane wave given by A. Erdélyi¹⁷, namely (in our notation)

$$\exp[-ix \sin \theta + iy \cos \theta]$$

$$= -\frac{e^{i\eta} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} [U_n(z) V_n(z') + U_n(-z) V_n(-z')] dn, \quad (5.13)$$

may be regarded as the sum of the negative of (5.8) and a similar expression with ξ and η replaced by $-\xi$ and $-\eta$. In informal discussions with the writer, Prof. Erdélyi has pointed out that the work leading to our expression (5.5) for the field may be considerably shortened by starting with some known integral for the impressed field, such as (5.13) or a related result. One way of doing this is to take

$$\exp[-ix \sin \theta + iy \cos \theta] + S_1,$$

¹⁷ Proc. Roy. Soc. Edinburgh, **61**, pp. 61-70, 1941.

as given by the left hand side of (5.8), to be the impressed field in the equation

$$E = \text{impressed field} + \text{reflected field}$$

From the form of (5.8) and the discussion of expression (4.6) (given between equations (4.14) and (4.15)) we are led to assume the reflected field to be an outgoing wave of the form

$$-\frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z') a(n) dn$$

where $a(n)$ must be chosen so as to make E vanish on the surface of the cylinder. This gives $a(n) = V_n(z'_0)/W_n(z'_0)$ and leads directly to the expression (5.11) for E .

When the incident wave is vertically polarized, integrals for H may be obtained from the series of Section 4 in much the same manner as were the integrals for E . The analogues of the earlier results are

$$H = \exp(-ix \sin \theta + iy \cos \theta) - \frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_1} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z')' U_n(z'_0) dn / W_n(z'_0), \quad (5.14)$$

$$H = \exp(-ix \sin \theta + iy \cos \theta) + S_1 + S_3(h), \quad (5.15)$$

$$S_3(h) = \frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z')' V_n(z'_0) dn / W_n(z'_0), \quad (5.16)$$

$$H = \frac{e^{iy} \sec \frac{\theta}{2}}{2i} \int_{L_2} \left(\frac{iw}{2}\right)^n \frac{\Gamma(n+1)}{\sin \pi n} U_n(z) W_n(z')' \left[\frac{V_n(z'_0)}{W_n(z'_0)} - \frac{V_n(z')}{W_n(z')} \right] dn, \quad (5.17)$$

$$H = -\pi e^{iy} \sec(\theta/2) \sum_{s=1}^{\infty} \left[\frac{(iw/2)^n \Gamma(n+1) U_n(z) W_n(z')' V_n(z'_0)}{\sin \pi n \partial' W_n(z'_0) / \partial n} \right]_{n=n'_s} \quad (5.18)$$

In these formulas $U_n(z)$, etc. are defined by (4.19); w, z, z' by (4.13), z'_0 by (4.17). In (5.14) w is restricted to $0 < w < 1$. In (5.15) $S_3(h)$ is given by (5.16), and w may be anywhere in $0 < w < \infty$. In (5.17) w

may also lie anywhere in $0 < w < \infty$, but in (5.18) it is restricted to $1 < w < \infty$ except when $\xi > \eta$ (roughly the shadow region) in which case w may be unity. In (5.18) $n = n'_s$ is the s th zero of $W_n(z'_0)$. The zeros of $W_n(z'_0)$ interlace those of $W_n(z'_0)$ shown in Fig. 5.1.

When $h = \eta_0^2/2 = 0$ the parabolic cylinder degenerates into a half-plane and our solutions reduce to Sommerfeld's expressions for waves diffracted by a half-plane. It may be shown that if

$$T_2 = \eta \cos \frac{\theta}{2} + \xi \sin \frac{\theta}{2} = (2r)^{1/2} \sin \left(\frac{\varphi + \theta}{2} + \frac{\pi}{4} \right), \quad (5.19)$$

we have

$$\begin{aligned} S_2(0) &= -(i/\pi)^{1/2} \exp[ix \sin \theta + iy \cos \theta] \int_{T_2}^{\infty} e^{-t^2} dt, \\ S_3(0) &= -S_2(0). \end{aligned} \quad (5.20)$$

When this expression for $S_2(0)$ is added to (5.6) we obtain Sommerfeld's result for the case of horizontal polarization.

One may verify that the series (5.12) leads to Sommerfeld's result as z'_0 approaches zero. By neglecting $O(z^2)$ terms in (9.3) and setting $n_s = -2s + p_s$ for the s th zero of $W_n(z'_0)$ we may obtain the following relations which are needed in the course of the verification

$$\begin{aligned} p_s &= -4iz'_0 \Gamma(s + 1/2) / \pi \Gamma(s) + \dots, \\ \partial W_n(z'_0) / \partial n \text{ at } n_s &= \Gamma(s)/4 + \dots, \\ W_n(z'_0) \text{ at } n_s &= -2iz'_0 \Gamma(s + 1/2) / \pi + \dots, \\ \left[\frac{V_n(z'_0)}{\sin \pi n \partial W_n(z'_0) / \partial n} \right]_{n=n_s} &= 2/\pi + \dots. \end{aligned} \quad (5.21)$$

6. SURFACE CURRENTS ON THE CYLINDER

As shown in Fig. 6.1, the surface current J on the perfectly conducting cylinder $\eta = \eta_0$ is parallel to the crest of the cylinder (and to the electric intensity E) when the incident wave is horizontally polarized. We have from Maxwell's equations in parabolic coordinates

$$J = [-H_\xi]_{\eta=\eta_0} = (i\zeta_0)^{-1} (2r)^{-1/2} [\partial E / \partial \eta]_{\eta=\eta_0}. \quad (6.1)$$

Here H_ξ is the component of magnetic intensity in the ξ -direction. ζ_0 is the intrinsic impedance of free space given by $\zeta_0 = (\mu_0/\epsilon_0)^{1/2} = \omega\mu_0$ where the second part of the equation follows from $2\pi/\lambda = \omega(\mu_0\epsilon_0)^{1/2}$

and $\lambda = 2\pi$. The E_0 and the H_0 of the incident wave are related by $H_0 = E_0/\zeta_0 = 1/\zeta_0$ since $E_0 = 1$. In rational MKS units $\zeta_0 = 120\pi$ ohms.

The derivative in (6.1) may be obtained by differentiating expression (5.11) for E and then setting $\eta = \eta_0$. Use of the Wronskian (9.9) for $V_n(z'_0)$, $W_n(z'_0)$ then takes (6.1) into

$$\zeta_0 J = M_0 \int_{L_2} \frac{dn}{\sin \pi n} (iw)^n U_n(z)/W_n(z'_0) \quad (6.2)$$

where $w = \tan(\theta/2)$, L_2 is shown in Fig. 5.1, z and z'_0 are given by (4.13) and (4.17), and

$$M_0 = (i/8\pi r)^{1/2} e^{-ir} \sec \frac{\theta}{2}, \quad (6.3)$$

$$r = (\xi^2 + \eta_0^2)/2.$$

In this Section r will be restricted to mean a radius vector drawn to the trace of the cylinder on the (x, y) plane of Fig. 6.1.

Closing L_2 on the right and on the left gives the two series

$$\zeta_0 J = 2iM_0 \sum_{n=0}^{\infty} (-iw)^n U_n(z)/W_n(z'_0), \quad 0 < w < 1, \quad (6.4)$$

$$\zeta_0 J = -2i\pi M_0 \sum_{s=1}^{\infty} \left[\frac{(iw)^n U_n(z)}{\sin \pi n \partial W_n(z'_0)/\partial n} \right]_{n=n_s}, \quad 1 < w < \infty, \quad (6.5)$$

where n_s is the s th zero of $W_n(z'_0)$ regarded as a function of n .

For the half-plane case $\eta_0 = 0$, and (9.4) gives

$$W_n(0) = -i^n/2\Gamma(1 + n/2).$$

In this case the series (6.4) may be expressed as an infinite integral

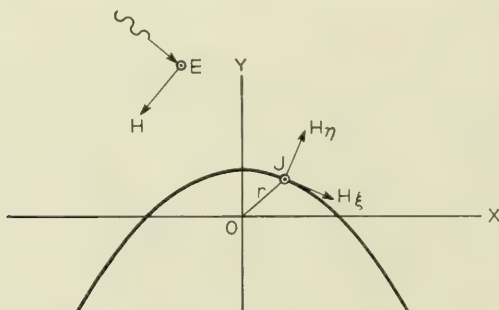


Fig. 6.1 — Relationship between surface current density J and electromagnetic field when incident wave is horizontally polarized. E and J are normal to the plane of the paper.

when the integral for $\Gamma(1 + n/2)$ is inserted and the sum (9.22) [i.e., the sum for the generating function of $U_n(z)$] used. Integrating part of the result gives

$$\begin{aligned} \xi_0 J = (2/i\pi r)^{1/2} \cos \frac{\theta}{2} \left[e^{-ir} \right. \\ \left. - 2i\xi \sin \frac{\theta}{2} e^{-ir \cos \theta} \int_{\xi \sin(\theta/2)}^{\infty} \exp(-it^2) dt \right]. \end{aligned} \quad (6.6)$$

In (6.6) r is the distance along the half-plane as measured from the edge: $r = \xi^2/2 = |y|$. Positive values of ξ correspond to the shadowed side of the half-plane and negative values to the illuminated side. With this interpretation (6.6) agrees with the current density obtained from Sommerfeld's expression for the field.

Although (6.6) has been derived from (6.2) on the assumption that $0 < w < 1$ it also holds for $0 < w < \infty$ as may be shown by analytic continuation. Again, (6.6) may be obtained from (6.5).

Since the factor $r^{-1/2}$ comes from the multiplier M_0 in (6.4), it is possible that (6.6) may give one an idea of how the current density behaves near the crest of a thin cylinder which is almost, but not quite, a half-plane. Of course, r would have to be interpreted as shown in Fig. 6.1.

In order to study J when the radius of curvature of the cylinder is large compared to a wavelength we consider the case $\theta = \pi/2$, i.e. $w = 1$, in which the incident wave comes in horizontally. In this case most of the variation of the current density occurs near the crest of the cylinder where, as it turns out, ξ is of the order of $\eta_0^{1/3}$, η_0 being large.

At the beginning of the investigation rough calculations of the integrand in (6.2), based on the asymptotic expressions of Section 12, suggested that for small ξ and large η_0 :

(a) Most of the contribution to the integral (6.2) comes from the neighborhood around point C shown on Fig. 6.2 where $m = n + 1 = z_0'^2/2 = -i\eta_0^2/2 = -ih$. Point C is a critical point associated with the asymptotic behavior of $W_n(z_0')$.

(b) The path of steepest descent for (6.2) roughly corresponds to the line ACD of Fig. 6.2, C being the high point of the path. Along this path $\text{Im}[f(t_0) - f(t_1)] = 0$ where $f(t) = -t^2 + 2zt - m \log t$, $m = n + 1$, is the function entering the definition (10.1) of the parabolic cylinder functions, and t_0, t_1 are the saddle points of $\exp[f(t)]$. This path in the n -plane separates the regions in which $W_n(z_0')$ has different asymptotic forms. It is the boundary of region III' in Fig. 12.2 and has been studied in Sections 11 and 12.

Once (b) is verified the truth of (a) follows almost immediately since

the path of integration L_2 may be deformed into ACD without passing over any singularities of the integrand of (6.2).

In order to verify (b), we note that the entries in Table 12.3 for $W_n(z'_0)$ show that along ACD

$$W_n(z'_0) \sim A'_0 \approx \exp [f(t_0)]. \quad (6.7)$$

Here the expressions for $W_n(z'_0)$ along ACD are taken to be those corresponding to the regions $I'b$ and II' shown in Fig. 12.2. In making the last approximation in (6.7) we have neglected the slowly varying coefficient of the exponential function in the expression (12.9) for A'_0 . Since $|\xi| \ll \eta_0$ we set $\xi = 0$ in $U_n(z)$. Then upon using the values (9.4) for $U_n(0)$, (6.7) for $W_n(z'_0)$, and unity for w , we see that the integrand of (6.2) behaves like

$$\frac{i^n \exp [-f(t_0)]}{2 \sin (\pi n/2) \Gamma(1 + n/2)}. \quad (6.8)$$

On ACD we have, in dealing with $W_n(z'_0)$, $-3\pi/2 < \arg m \leq -\pi/2$. Hence, from $t_0 + t_1 = i^{-1/2} \eta_0$ and from $f(t_0) + f(t_1)$ as calculated from (12.9), we have

$$\begin{aligned} \exp [f(t_0)] &= \exp \left(\frac{1}{2}[f(t_0) + f(t_1)] + \frac{1}{2}[f(t_0) - f(t_1)] \right) \\ &\sim i^{n+1} (2\pi)^{-1/2} \Gamma(-n/2) \exp \left(-\frac{i\eta_0^2}{2} + \frac{1}{2}[f(t_0) - f(t_1)] \right) \end{aligned} \quad (6.9)$$

where we have used the second of expressions (12.10) to evaluate $\exp [m(1 - \log (m/2))/2]$. Substitution of (6.9) in (6.8) shows that the integrand behaves roughly like

$$\exp \left(\frac{i\eta_0^2}{2} - \frac{1}{2}[f(t_0) - f(t_1)] \right). \quad (6.10)$$

The truth of statement (b) then follows from the fact that the lines of steepest descent of (6.10) in the n -plane are given by $Im [f(t_0) - f(t_1)] = 0$. To see that C is the high point of ACD we use (12.9) to show that near C we have

$$f(t_0) - f(t_1) \approx (2/3z'_0) (z_0'^2 - 2m)^{3/2}$$

where $m = n + 1$. Consequently, $f(t_0) - f(t_1)$ is real and positive on AC [where, near C , $\arg (z_0'^2 - 2m) = -\pi/6$] and on CD [where $\arg (z_0'^2 - 2m) = -3\pi/2$, m being in region II' according to the convention used in (6.7)]. That C is the high point now follows from (6.10).

In accordance with statement (a), we must study the form assumed by the integrand of (6.2) when n is near point C .

When (1) n is near C , (2) η_0 is large, and (3) $|\xi| \ll \eta_0$ we have for the various terms in (6.2)

$$i^n \sin \pi n \sim 2i^{1-n} \quad (6.11)$$

$$\Gamma(1 + n/2) W_n(z'_0) \sim (\eta_0/4)^{1/3} (2\pi)^{1/2} i^{5/6} \exp[-i\eta_0^2/2] Ai(\alpha),$$

$$2\Gamma(1 + n/2) U_n(z) \sim i^n \exp \left[\frac{i\xi^2}{2} + \xi \left(\frac{2m}{i} \right)^{1/2} - \frac{\xi^3}{6} \left(\frac{i}{2m} \right)^{1/2} \right] \quad (6.12)$$

where $Ai(\alpha)$ denotes the Airy integral defined by (13.12), $m = n + 1$, $\arg m$ is near $-\pi/2$, and

$$\alpha = (2/i\eta_0^2)^{1/3} (m + i\eta_0^2/2), \quad (6.13)$$

$$d\alpha = (2/i\eta_0^2)^{1/3} dn.$$

Expression (6.11) comes from (13.21) and expression (6.12) comes from region Ia of Table 12.2 (strictly speaking, region Ib should be used but point C is so close to $\arg m = -\pi/2$ that the simpler expression for Ia may be used). In obtaining (6.12) it is necessary to use the terms shown in the expansions (12.5) of t_0 and $\log t_1/t_0$. It may be shown that (6.12) also holds for negative values of ξ .

We now set $w = 1$ in (6.2) and change the variable of integration from n to α . Substituting for m in (6.12) its expression in terms of α , expanding in powers of α and neglecting higher order terms, converts the argument of the exponential function into

$$i\xi^2/2 - i\xi\eta_0 - i\gamma^3/3 + \alpha\gamma i^{1/3} \quad (6.14)$$

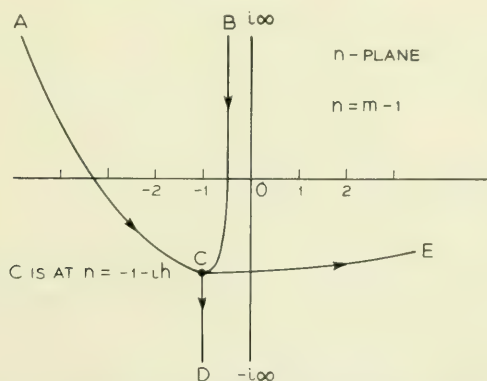


Fig. 6.2 — Paths of integration used in studying the current density and diffraction pattern when h is large. Path BCD is equivalent to path L_2 of Fig. 5.1. AC and CD are boundary lines which mark a change in the asymptotic behavior of $W_n(z'_0)$. Far out towards A the line AC tends to become parallel to BC .

where

$$\gamma = \xi/(2\eta_0)^{1/3} = x/2h^{2/3} \quad (6.15)$$

When our approximations are set in (6.2) we obtain

$$\zeta_0 J \approx (i/2\eta_0)^{2/3} \pi^{-1} \exp[-i\xi\eta_0 - i\gamma^3/3] \int_{-\infty \exp(i2\pi/3)}^{\infty \exp(-i2\pi/3)} \exp(i^{1/3}\gamma\alpha) d\alpha / Ai(\alpha). \quad (6.16)$$

Here we have taken the path of integration in the complex α -plane to be the transformed version of the path of steepest descent in the n -plane. That the path for (6.16) is still the one of steepest descent when $\gamma = 0$ and α is large follows from the asymptotic expansion (13.19) for $Ai(\alpha)$. It is interesting to note that near the crest of the cylinder $\xi\eta_0 + \gamma^3/3$ is approximately the distance along the cylinder as measured from the crest.

The expression (2.14) for $\zeta_0 J$ is obtained from (6.16) when $Ai(\alpha)$ is transformed by the relations (13.17).

When γ is positive the path of integration for α in (6.16) may be closed on the left to obtain a convergent series, the terms of which arise from the zeros of $Ai(\alpha)$. These zeros lie on the negative real α -axis starting with $\alpha = a_1 = -2.338 \dots$, $a_2 = -4.087 \dots$. The first fifty values of a_s and the corresponding values of the derivative $Ai'(a_s)$ have been tabulated*. Thus, when γ is positive,

$$\zeta_0 J \approx (2/i\eta_0^2)^{1/3} e^{-i\xi\eta_0 - i\gamma^3/3} \sum_{s=1}^{\infty} \frac{\exp(i^{1/3}\gamma a_s)}{Ai'(a_s)} \quad (6.17)$$

The leading term in this series leads to the approximation (2.19) when we use $Ai'(a_1) = .701 \dots$. Expression (6.17) gives the form assumed by (6.5) when $w = 1$ and $h \rightarrow \infty$. For large values of s^*

$$a_s \sim -[3\pi(4s - 1)/8]^{2/3} \\ Ai'(a_s) \sim -(-)^s \pi^{-1/2} (-a_s)^{1/4} \quad (6.18)$$

When γ is large and negative an asymptotic expansion for $\zeta_0 J$ may be obtained by setting the asymptotic expansion (13.19) in (6.16) and using the method of steepest descent. It is found that the saddle point is at $\alpha = \alpha_0 = \gamma^2 i^{2/3}$ and the slope of the path of the steepest descent through it is given by $\arg(d\alpha) = -5\pi/12$. This leads to the expression for $\zeta_0 J$ which appears in (2.17).

When the incident wave is vertically polarized the magnetic intensity

* Reference 11, page 424.

H is parallel to the crest of the parabolic cylinder. Since the cylinder is a perfect conductor, the current density J_r on the surface $\eta = \eta_0$ is equal in magnitude to H and its direction is that of increasing ξ . Thus setting $z' = z'_0$ in expression (5.17) for H and using the Wronskian (9.9) gives

$$J_r = N \int_{L_2} \frac{dn}{\sin \pi n} (iw)^n U_n(z), {}'W_n(z'_0), \quad (6.19)$$

$$N = e^{-ir} \left(\sec \frac{\theta}{2} \right) / 2\pi^{1/2}, \quad r = (\eta_0^2 + \xi^2)^{1/2},$$

where $'W_n(z'_0)$ is defined by (4.19).

Closing L_2 on the right and left leads to the analogues of (6.4) and (6.5):

$$J_r = 2iN \sum_{n=0}^{\infty} (-iw)^n U_n(z) {}'W_n(z'_0), \quad 0 < w < 1, \quad (6.20)$$

$$J_v = -2i\pi N \sum_{s=1}^{\infty} \left[\frac{(iw)^n U_n(z)}{\sin \pi n \partial {}'W_n(z'_0) / \partial n} \right]_{n=n'_s}, \quad 1 < w < \infty, \quad (6.21)$$

where n'_s is the s th zero of $'W_n(z'_0)$. The zeros of both $'W_n(z'_0)$ and $W_n(z'_0)$ are enclosed by the path of integration L_4 shown in Fig. 5.1.

The current density on a half-plane is obtained by setting $z'_0 = 0$ in (6.20), using $'W_n(0) = W_n'(0) = -i^{n-1}/\Gamma(n/2 + 1/2)$, from (9.4), and the generating function series (9.22) for $U_n(z)$:

$$J_v = 2(i/\pi)^{1/2} e^{-ir \cos \theta} \int_{\xi \sin(\theta/2)}^{\infty} e^{-it^2} dt \quad (6.22)$$

where r has the same significance as in (6.6). This agrees with the expression obtained from Sommerfeld's result for the half-plane.

When $w = 1$ and h is large, the path of steepest descent for (6.19) becomes the same as that for the case of horizontal polarization, namely ACD of Fig. 6.2. This follows from the fact that, as may be seen from (12.2), the controlling exponential functions for $'W_n(z'_0)$ and $W_n(z'_0)$ are the same. The analogue of (6.16) is obtained by using the approximation (13.24) for $'W_n(z'_0)$:

$$J_v \approx (1/2\pi i) \exp[-i\xi\eta_0 - i\gamma^3/3] \int_{\infty \exp(-i2\pi/3)}^{\infty \exp(-i2\pi/3)} \exp(i^{1/3}\gamma\alpha) d\alpha {}'Ai'(\alpha) \quad (6.23)$$

where $Ai'(\alpha)$ denotes $dAi(\alpha)/d\alpha$ and γ is given by (6.15).

For positive values of γ (6.23) and $d^2 Ai(\alpha)/d^2 \alpha = \alpha Ai(\alpha)$ lead to

$$J_v \approx -\exp[-i\xi\eta_0 - i\gamma^3/3] \sum_{s=1}^{\infty} \frac{\exp(i^{1/3}\gamma a'_s)}{a'_s Ai(a'_s)} \quad (6.24)$$

where $a'_1 = -1.019$, $a'_2 = -3.248$, $\dots$, etc. are the zeros of $Ai'(\alpha)$.* When we use $Ai(a'_1) = 0.5357$ the leading term in (6.24) gives (2.19). For large values of s^*

$$a'_s \sim -[3\pi(4s-3)/8]^{2/3},$$

$$Ai(a'_s) \sim -(-)^s (-a'_s)^{-1/4} \pi^{-1/2}.$$

The expression (2.17) for J_v when γ is large and positive may be obtained by applying the method of steepest descent to (6.23). The asymptotic expression for $Ai'(\alpha)$ is obtained by differentiating (13.19).

When the cylinder is a good conductor, but not perfect, the expression for H analogous to (4.25) leads to an integral for J_v , obtained from (6.23) by substituting $Ai'(\alpha) + \ell Ai(\alpha)$ for $Ai'(\alpha)$, which is equivalent to one given by Fock.† Here $\ell = -(ih)^{1/3} \zeta/\zeta_0$ is assumed to be small compared to unity and ζ/ζ_0 is the ratio of the intrinsic impedance of the cylinder material to that of free space. Horizontal incidence, $\theta = \pi/2$, is assumed.

The analogue of (4.25) for H has the same form as (4.21) except that now $'U_n(z'_0)$ is replaced by $'U_n(z'_0) + \tau U_n(z'_0)$, etc. The development leading from (4.21) through (5.15), (5.17), (6.19) to (6.23) may be carried out just as before. The work is also related to that given at the end of Section 7 where the effect of finite conductivity on the diffracted wave is briefly discussed.

A series corresponding to (6.24) may be derived from the integral. The exponential terms in this series are approximately $\exp[i^{1/3}\gamma(a'_s - \ell/a'_s)]$, and are similar to those in (7.63). Since $\zeta_0 = (\mu_0/\epsilon_0)^{1/2}$ is real and $\zeta \approx (i\omega\mu/g)^{1/2}$ when $g \gg \omega\epsilon$ (the notation is explained in connection with (4.24); the g denoting conductivity should not be confused with the g defined by (7.20)) the quantity $-i^{1/3}\ell/a'_s$ has a positive part. Thus, the attenuation of J_v in the shadow is decreased slightly when the conductivity g of the cylinder is reduced from infinity to a large finite value.

7. FIELD AT A GREAT DISTANCE BEHIND THE PARABOLIC CYLINDER

The field at any point, for the case of horizontal polarization, is given by expression (5.5) with $S_2(h)$ given by (5.4). Since $S_2(h)$ is the

* Reference 11, page 424.

† Reference 5, page 418.

only troublesome term in (5.5) most of this section will be devoted to its study. Far behind the cylinder ξ and η are large and positive, and the corresponding terms in the integrand of (5.4) are

$$\Gamma(n+1)U_n(z)W_n(z') = (2i\xi/\eta)^n i^{3/2} e^{-i\eta^2} (1+f)/2\eta\pi^{1/2}, \quad (7.1)$$

where the asymptotic expressions (9.16) and (9.17) give

$$1+f = 1 + \frac{in(n-1)}{4\xi^2} + \frac{i(n+1)(n+2)}{4\eta^2} + O(n^{1/2}r^2). \quad (7.2)$$

In writing the "order of" term it is assumed that ξ and η are both $O(r^{1/2})$ with $r \gg n^2$.

When (7.1) is put in (5.4) we obtain

$$S_2(h) = M_1 \int_{L_2} \frac{\beta^n i^{2n} V_n(z'_0)(1+f)}{\sin \pi n W_n(z'_0)} dn \quad (7.3)$$

where, from the expressions (4.1) for ξ and η ,

$$M_1 = [(i/\pi)^{1/2} e^{-ir} \sec(\theta/2)]/4\eta, \quad (7.4)$$

$$\beta = \xi w'/\eta = \xi[\tan(\theta/2)]/\eta = \cot(\varphi/2 + \pi/4) \tan(\theta/2).$$

Although it is not proved here, there is good reason to believe that (7.3) can be written as

$$S_2(h) = M_1 \int_{L_2} \frac{i^{2n} \beta^n V_n(z'_0)}{\sin \pi n W_n(z'_0)} dn + O(h^3 r^{3/2}) \quad (7.5)$$

when r becomes large and we restrict ourselves to the region $|\varphi| < \pi/2$ in order to get $O(\xi^2) = O(\eta^2) = O(r)$. The first term contains r only through the factor M_1 and is of order $r^{-1/2}$. The "order of" term assumes h to be moderately large compared to unity but $h^2 \ll r$. When $h < 1$ the h^3 is to be replaced by unity.

The general idea leading to (7.5) is that (7.2) may be used over the portion of L_2 where the integrand is large and important. On the portion where (7.2) differs appreciably from unity the integrand is negligibly small. The important portion of L_2 runs from $n = 1/2$ to $n = -1/2 - ih$ (approximately). In particular the variation of $i^{2n} V_n(z'_0)/\sin \pi n W_n(z'_0)$ along L_2 may be summarized as follows: from $-1/2$ to $+i\infty$ it decreases exponentially as i^{2n} , from $-1/2$ to $-ih$ it is equal to $-2i$ plus an oscillating function of order unity, and from $-ih$ to $-i\infty$ it decreases slowly at first and then more rapidly until it goes down like i^{-2n} (steepest descent behavior). This may be shown with the help of Fig. 12.2, the entries for regions $I'a$ and II' in Table 12.3, and the following items [see (12.9) and Figs. 10.1 and 10.2 for $z = i^{-1/2} \eta_0$ with $-3\pi/2 \leq \arg(i\eta_0^2$

$-2m) < \pi/2]$:

(a) $Re [f(t_1) - f(t_0)]$ is almost zero between $-\frac{1}{2}$ and $-ih$.

(b) $Im [f(t_1) - f(t_0)]$ is almost zero between $-ih$ and $-i\infty$.

(c) t_1/t_0 runs from zero to unity as n goes from -1 to $-1-ih$.

Items (a) and (b) are consistent with $f(t_0) - f(t_1) \approx (z'_0/2)(z_0'^2 - 2m)^{3/2}$ which holds when n is near $-ih$ and which was mentioned in connection with (6.10).

This concludes our discussion of the reasons for believing that (7.5) is true for general values of h . Now we shall check it for the special case $h = 0$.

When we set $h = 0$ (i.e. $z'_0 = 0$) in (7.3), use (9.4) and close L_2 by an infinite semicircle, we obtain

$$S_2(0) \sim \frac{i^{3/2} e^{-ir}}{\pi^{1/2} 2T_2} \left[1 - \frac{1}{2iT_2^2} + \cdots \right]. \quad (7.6)$$

This agrees with the first two terms in the asymptotic expansion of the Fresnel integral expression (5.20) for $S_2(0)$. In expanding (5.20) we need the first of the two asymptotic expansions (both of which hold for $T \gg 1$)

$$\int_T^\infty e^{-it^2} dt \sim -\frac{i \exp(-iT^2)}{2T} \left[1 - \frac{1}{2iT^2} + \cdots \right], \quad (7.7)$$

$$\int_{-\infty}^T e^{-it^2} dt \sim (\pi/i)^{1/2} + \frac{i \exp(-iT^2)}{2T} \left[1 - \frac{1}{2iT^2} + \cdots \right], \quad (7.8)$$

and also the first of the relations

$$\begin{aligned} x \sin \theta + y \cos \theta - T_2^2 &= -r, \\ T_2 &= \eta(1 + \beta) \cos(\theta/2). \end{aligned} \quad (7.9)$$

In much of the following work we shall assume ξ and η to be so great that we can neglect the terms denoted by $O(h^3/r^{3/2})$ in (7.5). We shall use the asymptotic sign $\sim$ to acknowledge this omission.

From (7.4), β is equal to unity when $\varphi = \theta - \pi/2$. When r is very large this value of φ marks the shadow boundary. In the shadow $\beta > 1$ and in the illuminated region $\beta < 1$.

Closing L_2 on the right and on the left converts (7.5) into

$$S_2(h) \sim 2iM_1 \sum_{n=0}^{\infty} \beta^n V_n(z'_0)/W_n(z'_0), \quad (7.10)$$

$$S_2(h) \sim 2iM_1 \left\{ (\beta - 1)^{-1} - \pi \sum_{s=1}^{\infty} \left[\frac{i^{2n} \beta^n V_n(z'_0)}{\sin \pi n \partial W_n(z'_0)/\partial n} \right]_{n=n_s} \right\} \quad (7.11)$$

It can be shown that (7.10) converges if $\beta < 1$ (see (4.18)) and that (7.11) converges if $\beta > 1$ (see (12.13)). The term $1/(\beta - 1)$ in (7.11) comes from the poles of $\csc \pi n$ inside the path L_3 shown in Fig. 5.1. From (7.7) and

$$\begin{aligned} -x \sin \theta + y \cos \theta - T_1^2 &= r, \\ T_1 &= \eta(1 + \beta) \cos(\theta/2) \end{aligned} \quad (7.12)$$

it may be shown that when $\beta > 1$, so that T_1 is negative, (5.6) has the asymptotic expression

$$\exp[-ix \sin \theta + iy \cos \theta] + S_1 \sim 2iM_1/(1 - \beta). \quad (7.13)$$

When this is added to (7.11) the $1/(1 - \beta)$ terms cancel leaving a series for E valid in the shadow where $\beta > 1$:

$$E \sim -2iM_1\pi \sum_{s=1}^{\infty} \left[\frac{i^{2n} \beta^n V_n(z'_0)}{\sin \pi n \partial W_n(z'_0)/\partial n} \right]_{n=n_s} \quad (7.14)$$

This series may also be obtained from the more general series (5.12) for E by using (7.1) and neglecting f .

We now take up the problem of finding the paths of steepest descent for the integral in (7.5) when β is near unity and h is large. When $\beta = 1$ and h is large, the integrand in (7.5) may be expressed in terms of $\exp[f(t_1) - f(t_0)]$ by using Table 12.3. In Section 6 it has been pointed out that the path of steepest descent for $\exp[f(t_1) - f(t_0)]$ is the path ACD of Fig. 6.2, with C being the high point. This suggests that the path ACD should be used to deal with the terms in (7.5) leading to $\exp[f(t_1) - f(t_0)]$. These terms are $U_n(z'_0)/W_n(z'_0)$ (introduced by the use of (4.8)) for the portion of L_2 between B and C , and $V_n(z'_0)/W_n(z'_0)$ for the portion between C and D . As a further argument supporting the use of the path AC we note that when n is on AC , i.e., on the edge of region $I'b$, Table 12.3 gives

$$U_n(z'_0)/W_n(z'_0) \sim -i(1 - i^{-4n})(t_1/t_0)^{1/2} \exp[f(t_1) - f(t_0)]. \quad (7.15)$$

Hence the variation of $i^{2n} \csc \pi n$ in the integrand of (7.5) is just cancelled by that of $(1 - i^{-4n})$ in (7.15). Consequently $i^{2n} U_n(z'_0)/[\sin \pi n W_n(z'_0)]$ varies as $\exp[f(t_1) - f(t_0)]$ along AC (the variation of t_1/t_0 is relatively small).

These considerations lead us to write (5.4) as

$$\begin{aligned}
S_2(h) &= S_{21} + S_{22} + S_{23} \\
S_{21} &= - \int_B^C F \, dn, \\
S_{22} &= - \int_A^C F U_n(z'_0) \, dn / W_n(z'_0), \\
S_{23} &= \int_C^D F V_n(z'_0) \, dn / W_n(z'_0), \\
F &= e^{iy} \sec(\theta/2) (iw/2)^n \Gamma(n+1) U_n(z) W_n(z') / 2i \sin \pi n.
\end{aligned} \tag{7.16}$$

When instead of (5.4) the expression (7.5) for $S_2(h)$ is used we obtain

$$\begin{aligned}
S_{21} &\sim -M_1 \int_B^C i^{2n} \beta^n \, dn / \sin \pi n, \\
S_{22} &\sim -M_1 \int_A^C i^{2n} \beta^n U_n(z'_0) \, dn / [\sin \pi n W_n(z'_0)], \\
S_{23} &\sim M_1 \int_C^D i^{2n} \beta^n V_n(z'_0) \, dn / [\sin \pi n W_n(z'_0)].
\end{aligned} \tag{7.17}$$

In S_{22} it is permissible to swing AC from its original position BC because the zeros of $U_n(z'_0)$ cancel those of $\sin \pi n$. When $\beta = 1$, AC and CD are the paths of steepest descent for S_{22} and S_{23} in (7.17) because $\text{Im}[f(t_1) - f(t_0)] = 0$ on ACD .

The asymptotic expression (7.17) for S_{21} may be evaluated by temporarily assuming β to be a complex number with $|\beta| < 1$ and $|\arg \beta| < \pi/2$. The integral along BC is the integral along BCE minus the integral along CE (see Fig. 6.2). Deforming BCE into L_1 of Fig. 5.1 shows that its contribution to S_{21} is $-2iM_1/(1 - \beta)$. An infinite series for the integral along CE may be obtained by expanding $i^{2n}/\sin \pi n$ in powers of $\exp(-i2\pi n)$ and integrating termwise from $n = n_0 = -1 - ih$ to $n = \infty - ih$, i.e., from C to E . In this way we obtain

$$S_{21} \sim -2iM_1 \left[\frac{1}{1 - \beta} + \sum_{t=0}^{\infty} \frac{\exp(n_0 \log \beta - 2\pi th)}{\log \beta - 2\pi it} \right]. \tag{7.18}$$

Despite the appearance of the right hand side, it is analytic around $\beta = 1$ and analytic continuation may be used to show that (7.18) holds for $0 < \beta < \infty$.

When h is large only the first term in the series is important and we have

$$\begin{aligned}
S_{21} &\sim 2iM_1[(\beta - 1)^{-1} - \beta^{-1-ih}/\log \beta] \\
&= 2iM_1(\beta - 1)^{-1} + iM_2 g^{-1}
\end{aligned} \tag{7.19}$$

where we have introduced two quantities which will be used later:

$$M_2 = 2M_1 h^{1/3} \beta^{-1-ih} = \left(\frac{i}{\pi}\right)^{1/2} \frac{h^{1/3} \exp[-ir + i g h^{2/3}]}{2\xi \sin(\theta/2)}, \quad (7.20)$$

$$g = -h^{1/3} \log \beta.$$

When $\beta = 1$

$$S_{21} \sim 2iM_1(ih + 1/2). \quad (7.21)$$

When h is large most of the contributions to the integrals (7.17) for S_{22} and S_{23} come from around $n = n_0 = -1 - ih$, and we may use the approximations

$$U_n(z'_0)/W_n(z'_0) \sim i^{-4/3} Ai(\alpha i^{-4/3})/Ai(\alpha), \quad (7.22)$$

$$V_n(z'_0)/W_n(z'_0) \sim i^{4/3} Ai(\alpha i^{4/3})/Ai(\alpha), \quad (7.23)$$

$$\alpha = (ih)^{-1/3}(n + 1 + ih), \quad n - n_0 = \alpha(ih)^{1/3},$$

which come from (13.21). Setting these in the integrals (7.17) and using the fact that $i^{2n}/\sin \pi n$ is nearly $2i$ around n_0 leads to

$$S_{22} \sim -M_2 \int_{-\infty \exp(i2\pi/3)}^0 \exp(-i^{1/3}\alpha g) Ai(\alpha i^{-4/3}) d\alpha / Ai(\alpha), \quad (7.24)$$

$$S_{23} \sim -i^{2/3} M_2 \int_0^{\infty \exp(-i2\pi/3)} \exp(-i^{1/3}\alpha g) Ai(\alpha i^{4/3}) d\alpha / Ai(\alpha), \quad (7.25)$$

$$S_{22} + S_{23} \sim iM_2 \Psi(g), \quad (7.26)$$

where

$$\begin{aligned} \Psi(g) = & i \int_{-\infty i^{4/3}}^0 \exp(-i^{1/3}\alpha g) Ai(\alpha i^{-4/3}) d\alpha / Ai(\alpha) \\ & + i^{5/3} \int_0^{\infty i^{-4/3}} \exp(-i^{1/3}\alpha g) Ai(\alpha i^{4/3}) d\alpha / Ai(\alpha). \end{aligned} \quad (7.27)$$

The expression (2.11) for $\Psi(g)$ is obtained from (7.27) by changing the variables of integration and using the transformations (13.17) for $Ai(\alpha)$.

Thus, when h is large and β near unity, (7.19) and (7.26) give

$$S_2(h) \sim 2iM_1(\beta - 1)^{-1} + iM_2[g^{-1} + \Psi(g)]. \quad (7.28)$$

In the shadow, where $\beta > 1$ and g is negative, (7.13) and (7.28) give

$$\begin{aligned} E = & \exp[-ix \sin \theta + iy \cos \theta] + S_1 + S_2(h) \\ & \sim iM_2 [g^{-1} + \Psi(g)]. \end{aligned} \quad (7.29)$$

This and the series (7.14) for E suggest that $\Psi(g) + 1/g$ may be ex-

pressed as a series in which the parabolic cylinder functions in (7.14) are replaced by Airy integrals. One way of obtaining this series from (7.14) is to use the Airy integral approximations (13.21). The zeros $n_1, n_2, \dots$ of $W_n(z'_0)$ go into the zeros $a_1, a_2, \dots$ of $Ai(\alpha)$ by virtue of the relation $n - n_0 = \alpha(ih)^{1/3}$ and we have

$$E \sim i^{-2/3} M_2 \sum_{s=1}^{\infty} \frac{\exp[-a_s g i^{1/3}]}{[Ai'(a_s)]^2}, \quad (7.30)$$

$$\Psi(g) + 1/g = -i^{1/3} \sum_{s=1}^{\infty} \frac{\exp[-a_s g i^{1/3}]}{[Ai'(a_s)]^2}, \quad (7.31)$$

where $g < 0$. Here, as in (6.17), $a_1 = -2.338 \dots$ and $Ai'(a_1) = .701 \dots$. In obtaining these relations we have used $i^{2n}/\sin \pi n \approx 2i$, and

$$Ai(a_s i^{4/3}) = -i^{5/3} Bi(a_s)/2 = i^{5/3}/2\pi Ai'(a_s), \quad (7.32)$$

where the first equation follows from (13.17) and the second from the Wronskian

$$Ai(\alpha)Bi'(\alpha) - Ai'(\alpha)Bi(\alpha) = 1/\pi. \quad (7.33)$$

The equal sign in (7.31) holds even though the steps leading to it indicate that $\sim$ should be used. This may be seen from an alternative derivation of (7.31) in which $Ai(\alpha i^{-4/3})$ in the first integral of (7.27) is replaced by the right hand side of*

$$Ai(\alpha i^{-4/3}) = -i^{4/3} Ai(\alpha) - i^{-4/3} Ai(\alpha i^{4/3}). \quad (7.34)$$

In the first portion the $Ai(\alpha)$'s cancel and the resulting integral contributes $-1/g$ to (7.27) (g must be negative for convergence). The second portion combines with the second integral in (7.27) to give a contour integral which leads to the series in (7.31) when the path of integration in the α -plane is closed on the left. The closure may be justified by the asymptotic expressions (13.19) and (13.20) for $Ai(\alpha)$ (again g must be negative).

Since the integrals in (7.27), and their equivalents in (2.11), converge uniformly for all finite values of g , $\Psi(g)$ is an integral function of g . When g is negative $\Psi(g)$ may be computed from the series (7.31). When g is positive I have not been able to find a practicable method of obtaining $\Psi(g)$ other than the numerical integration of (2.11). The results are shown in Table 2.1. Since $\Psi(g)$ is an integral function its Taylor's series about, say, $g = -.5$ converges for all values of g . The coefficients in this series may be computed from (7.31). However, I was unable to

* Reference 11, page 424.

obtain useful results by this method because the computation of the coefficients becomes more and more difficult.

When g is large and positive it may be shown that

$$\Psi(g) \sim -g^{-1} + (i\pi g)^{1/2} \exp(-ig^3/12). \quad (7.35)$$

The procedure used to establish (7.35) is much the same as that used to establish the more general result

$$\begin{aligned} S_{22} + S_{23} &\sim -iM_2g^{-1} \\ &- \left[\frac{h(1-\beta)}{r(1+\beta)} \right]^{1/2} \frac{\exp[-ir + 2ih(1-\beta)/(1+\beta)]}{\sin \frac{1}{2}(\varphi + \theta + \pi/2)}, \quad (7.36) \\ \frac{1-\beta}{1+\beta} &= \frac{\sin \frac{1}{2}(\varphi - \theta + \pi/2)}{\sin \frac{1}{2}(\varphi + \theta + \pi/2)}, \end{aligned}$$

which holds when $h^{1/3}(1-\beta)$ is large and positive.

When φ is near $-\pi/2 + \theta$, (7.36) gives the same result as (7.26) plus (7.35). For φ near $-\pi/2 + \theta$,

$$g \approx h^{1/3}(1-\beta) \approx [2h^{1/3}/\sin \theta] \sin \frac{1}{2}\left(\varphi - \theta + \frac{\pi}{2}\right), \quad (7.37)$$

which shows that g is proportional to the cube root of the radius of curvature $2h/\sin^3 \theta$ at the point where the incident ray is tangent to the cylinder.

When $\beta < 1$, (7.36) may be obtained from

$$S_{22} + S_{23} \sim -M_1 \int_C \frac{i^{2n} \beta^n}{\sin \pi n} dn - M_1 \int_{A'E} \frac{i^{2n} \beta^n U_n(z'_0)}{\sin \pi n W_n(z'_0)} dn. \quad (7.38)$$

The second integral in (7.38) represents, asymptotically, the wave reflected by the cylinder. This interpretation is suggested by the fact that, when the expression (7.1) for $\Gamma(n+1)U_n(z)W_n(z')$ is substituted in expression (5.1) for E , the resulting integral may be written as the second integral in (7.38).

The first term in (7.36) is obtained when $i^{2n}/\sin \pi n$ in the first integral of (7.38) is approximated by $2i$ and the result integrated. When the integrand of the second integral is examined with the help of Table 12.3, it is found to have a saddle point* at $m = m_1$ on the imaginary axis between $m = 0$ and $m = -ih$. Near m_1 the integrand is approximately

$$2\beta^{-1}(t_1/t_0)^{1/2} \exp[F(m)] \quad (7.39)$$

* It is interesting to note that a saddle point also appears in the study of reflection from a sphere. See page 86 of reference.⁷

where $F(m) = f(t_1) - f(t_0) + m \log \beta$ and $F'(m) = \log(t_0 \beta / t_1)$. Here t_0 and t_1 are functions of m defined by (12.9). At m_1 we have $t_0 \beta = t_1$ and this leads to

$$m_1 = -4ih\beta/(1 + \beta)^2, \quad F(m_1) = 2ih(1 - \beta)/(1 + \beta),$$

$$F''(m_1) = -(1 + \beta)/m_1(1 - \beta). \quad (7.40)$$

When we attempt to deform the path of integration ACE of the second integral in (7.38) into a path of steepest descent, we encounter no trouble near m_1 in regions $I'a$ and $I'b$. The path passes through m_1 with $\arg(dm) = -\pi/4$. Soon after passing through m_1 the path of steepest descent strikes the boundary between $I'a$ and II' at a point we shall call G . At G the imaginary part of m is $2h(1 - \beta)/(1 + \beta) \log \beta$. The asymptotic approximation to the integrand changes its form at this point. The choice of the path from G out to ∞ is not important since it contributes little to the value of the integral. However, if we insist on following paths of steepest descent, it turns out that we must split the path of integration at G .

When $h^{1/3}(1 - \beta) \gg 1$, it may be shown that the value of second integral in (7.38) is nearly equal to

$$-2M_1[-2\pi/\beta F''(m_1)]^{1/2} \exp[F(m_1)]$$

and this gives the second term in (7.36).

So far, in this section, we have been dealing with the case of horizontal polarization. Since the work for the case of vertical polarization (in which H plays the role of the wave function) follows much the same lines, we shall merely list the formulas corresponding to those already obtained for horizontal polarization. M_1 , β and M_2 , g are still given by (7.4) and (7.20), respectively.

$$S_3(h) = M_1 \int_{L_2} \frac{i^{2n} \beta^n {}'V_n(z'_0)}{\sin \pi n {}'W_n(z'_0)} dn + O(h^3/i^{3/2}), \quad (7.41)$$

$$S_3(0) = -S_2(0), \quad (7.42)$$

$$S_3(h) \sim 2iM_1 \sum_{n=0}^{\infty} \beta^n {}'V_n(z'_0) {}'W_n(z'_0), \quad \beta < 1 \quad (7.43)$$

$$S_3(h) \sim 2iM_1 \left\{ (\beta - 1)^{-1} - \pi \sum_{s=1}^{\infty} \left[\frac{i^{2n} \beta^n {}'V_n(z'_0)}{\sin \pi n \partial {}'W_n(z'_0)/\partial n} \right]_{n=n'_s} \right\}, \quad (7.44)$$

$\beta > 1$ (shadow region),

$n'_s = \text{sth zero of } {}'W_n(z'_0),$

$$H \sim -2iM_1\pi \sum_{s=1}^{\infty} \left[\frac{i^{2n}\beta^{n-1}V_n(z'_0)}{\sin \pi n \partial' W_n(z'_0) \partial n} \right]_{n=n'}, \quad \beta > 1 \quad (7.45)$$

$$S_3(h) = S_{31} + S_{32} + S_{33},$$

$$S_{31} = S_{21} \text{ defined by (7.16),} \quad (7.46)$$

$$S_{32} = S_{22} \text{ with } 'U_n(z'_0)'W_n(z'_0) \text{ in place of } U_n(z'_0)W_n(z'_0),$$

$$S_{33} = S_{23} \text{ with } 'V_n(z'_0)'W_n(z'_0) \text{ in place of } V_n(z'_0)W_n(z'_0),$$

$$S_{32} \sim i^{2/3}M_2 \int_{-\infty \exp(i2\pi/3)}^0 \exp(-i^{1/3}\alpha g) A i'(\alpha i^{-4/3}) d\alpha / A i'(\alpha), \quad (7.47)$$

$$S_{33} \sim M_2 \int_0^{\infty \exp(-i2\pi/3)} \exp(-i^{1/3}\alpha g) A i'(\alpha i^{4/3}) d\alpha / A i'(\alpha), \quad (7.48)$$

$$S_{32} + S_{33} \sim iM_2\Psi_r(g), \quad (7.49)$$

$$\begin{aligned} \Psi_r(g) = & i^{-1/3} \int_{-\infty \exp(i2\pi/3)}^0 \exp(-i^{1/3}\alpha g) A i'(\alpha i^{-4/3}) d\alpha / A i'(\alpha) \\ & - i \int_0^{\infty \exp(-i2\pi/3)} \exp(-i^{1/3}\alpha g) A i'(\alpha i^{4/3}) d\alpha / A i'(\alpha), \end{aligned} \quad (7.50)$$

$$H \sim iM_2[g^{-1} + \Psi_r(g)], \quad g < 0 \quad (7.51)$$

$$H \sim i^{-2/3}M_2 \sum_{s=1}^{\infty} \frac{\exp(-a'_s g i^{1/3})}{(-a'_s)[A i(a'_s)]^2}, \quad g < 0 \quad (7.52)$$

$$\Psi_r(g) + g^{-1} = -i^{1/3} \sum_{s=1}^{\infty} \frac{\exp(-a'_s g i^{1/3})}{(-a'_s)[A i(a'_s)]^2}, \quad g < 0 \quad (7.53)$$

$$a'_s = \text{sth zero of } A i'(\alpha), \quad a'_1 = -1.019, \quad A i(a'_1) = 0.5357,$$

$$A i'(a'_s i^{4/3}) = -i^{1/3} B i'(a'_s), \quad 2 = -i^{1/3} [2\pi A i(a'_s)], \quad (7.54)$$

$$A i''(\alpha) = \alpha A i(\alpha).$$

When g is large and positive,

$$\Psi_r(g) \sim -g^{-1} - (i\pi g)^{1/2} \exp(-ig^3/12), \quad (7.55)$$

$$\begin{aligned} S_{32} + S_{33} \sim & -iM_2 g^{-1} \\ & + \left[\frac{h(1-\beta)}{r(1+\beta)} \right]^{1/2} \frac{\exp[-ir + 2ih(1-\beta)(1+\beta)]}{\sin \frac{1}{2}(\varphi + \theta + \pi/2)}. \end{aligned} \quad (7.56)$$

The change in sign of the second term on the right in going from (7.36) to (7.56) comes from (12.2) and the analogous expression for $'W_n(z'_0)$

(only t_1 contributes to $U_n(z'_0)$ and only t_0 to $W_n(z'_0)$ at the saddle point m_1 of the second integral in (7.38)).

So far in this section the parabolic cylinder has been assumed to possess infinite conductivity. When the cylinder has a finite (but very large) conductivity, it may be shown that the field far out in the shadow is approximately

$$E \sim M_1 \int_{L_4} \frac{i^{2n} \beta^n}{\sin \pi n} \frac{V_n(z'_0) + \sigma^{-1} V_n(z'_0)}{W_n(z'_0) + \sigma^{-1} W_n(z'_0)} dn. \quad (7.57)$$

Equation (7.57) is suggested by (7.14) and (4.25). The analogue of (7.57) for vertical polarization may be obtained by replacing E , σ in (7.57) by H , τ so that $V_n(z'_0) + \tau V_n(z'_0)$ appears in place of $V_n(z'_0) + \sigma^{-1} V_n(z'_0)$, and so on.

When the parabolic cylinder functions are replaced by Airy integrals according to (13.21) and (13.24), equation (7.57) may be written as

$$E \sim i^{8/3} M_2 \int_{L_4} \frac{[\exp(-\alpha g i^{1/3})] Ai[(\alpha + k) i^{4/3}]}{Ai(\alpha + k)} d\alpha \quad (7.58)$$

where g and M_2 are given by (7.20), α by (7.23) and

$$k = -(ih)^{-1/3} \zeta / \zeta_0. \quad (7.59)$$

$|k|$ is small compared to unity. L_4 is a path of integration in the α plane which encloses the zeros of $Ai(\alpha + k)$ in a clockwise direction. Changing the variable of integration in (7.58) to $u = \alpha + k$ enables us to conclude that

$$\left[\begin{array}{c} E \text{ for finite} \\ \text{conductivity} \end{array} \right] \approx \left[\exp \left(-\frac{\zeta \psi}{\zeta_0} \right) \right] \left[\begin{array}{c} E \text{ for infinite} \\ \text{conductivity} \end{array} \right]. \quad (7.60)$$

Since we have assumed $\theta = \pi/2$, the relation (7.60) holds in the region where the angle ψ defined by Fig. 2.3 is negative.

The analogue of (7.58) for vertical polarization is obtained by replacing E by H , omitting the $i^{8/3}$, and replacing the ratio of the Airy integrals by

$$Ai'[(\alpha + \ell/\alpha) i^{4/3}] / Ai'(\alpha + \ell/\alpha) \quad (7.61)$$

where

$$\ell = -(ih)^{1/3} \zeta / \zeta_0. \quad (7.62)$$

Even though h is large, ζ/ζ_0 is assumed to be so small that ℓ is small compared to unity. The path of integration L'_4 must now enclose the zeros of $Ai'(\alpha + \ell/\alpha)$ which are close to those of $Ai'(\alpha)$ at $\alpha = a'_s, s = 1,$

2, $\dots$. It must not pass close to $\alpha = 0$ since the work leading to (7.61) assumes ℓ/α to be a small number. Changing the variable of integration to $v = \alpha + \ell/\alpha$, approximating ℓ/α , ℓ/α^2 by ℓ'_s/v , ℓ'_s/v^2 , and evaluating the integral by considering the residues of the poles at $v = a'_s$ gives

$$H \approx i^{-2/3} M_2 \sum_{s=1}^{\infty} \left(1 - \frac{\ell}{a_s'^2} \right)^{-1} \frac{\exp [-i^{1/3} g(a'_s - \ell'_s/a'_s)]}{(-a'_s)[4i(a'_s)]^2}. \quad (7.63)$$

This shows, to a first approximation, how the expression (7.52) is modified when the cylinder is a very good, but not perfect, conductor. Of course g must be negative in (7.63). Since ℓ in (7.62) varies as $h^{1/3}$ while k in (7.59) varies as $h^{-1/3}$ it appears that the field for vertical polarization is much more sensitive to changes in the conductivity than it is for horizontal polarization.

It may be verified that the change in the exponential terms in the series (7.30) and (7.52) produced by finite conductivity, namely

$$\begin{aligned} a_s &\text{ changes to } a_s - k \\ a'_s &\text{ changes to } a'_s - \ell/a'_s, \end{aligned} \quad (7.64)$$

agrees, to a first approximation, with the change produced in the corresponding series (given, for example, by the series (27) and (28) on page 45 of Reference 7) for the propagation of radio waves over the earth's surface.

8. FIELD AT A GREAT DISTANCE BEHIND THE PARABOLIC CYLINDER WHEN $\theta = \pi/2$ AND h IS LARGE

In the work of Section 7 the angle of incidence θ may lie anywhere between 0 and π . Here we take $\theta = \pi/2$, which corresponds to the case shown in Fig. 2.3 and described in Section 2. Some simplification is obtained thereby. For example, the incident wave is now simply $\exp(-ix)$. We shall write the expressions for the horizontal and vertical polarization cases as

$$E = (e^{-ix} + S_1)_r + S_{21} + (S_{22} + S_{23}), \quad (8.1)$$

$$H = (e^{-ix} + S_1)_r + S_{21} + (S_{32} + S_{33}), \quad (8.2)$$

respectively. Here S_{21} , $\dots$ are defined by (7.16) and (7.46) in which

$$\begin{aligned} \theta &= \pi/2, & w &= 1, \\ \beta &= \xi/\eta = \cot(\varphi/2 + \pi/4) = 1 - \varphi + \varphi^2/2 - \varphi^3/3 + \dots \end{aligned} \quad (8.3)$$

Throughout this section β will be defined by (8.3), i.e., by (7.4) with

$\theta = \pi/2$. Also, from (5.3) and (5.6)

$$(e^{-ix} + S_1)_r = (i/\pi)^{1/2} e^{-ix} \int_{-\infty}^{T_1} e^{-it^2} dt, \quad T_1 = 2^{-1/2}(\eta - \xi) \quad (8.4)$$

$$= (2r)^{1/2} \sin(\varphi/2).$$

The subscript r is used to denote correspondence to a half-plane with its edge at $r = 0$.

When h is large, physical reasons lead us to expect a similarity between our field and the one behind a half-plane with its edge at the crest of the cylinder where $\rho = 0$ (see Fig. 2.3). The main part of this field is the analogue of (8.4):

$$(e^{-ix} + S_1)_\rho = (i/\pi)^{1/2} e^{-ix} \int_{-\infty}^{T_3} e^{-it^2} dt, \quad T_3 = (2\rho)^{1/2} \sin(\psi/2), \quad (8.5)$$

where the subscript ρ indicates that $[\exp(-ix) + S_1]_\rho$ corresponds to diffraction behind the half-plane just mentioned.

In order to make use of the similarity between the field behind the cylinder and the half-plane with its edge at the crest of the cylinder, we change the polar coordinates from (r, φ) to (ρ, ψ) . From

$$\rho e^{i\psi} = r e^{i\varphi} - ih \quad (8.6)$$

it may be shown that, when h^2/r is small,

$$(e^{-ix} + S_1)_r - (e^{-ix} + S_1)_\rho = (i/\pi)^{1/2} e^{-ix} \int_{T_3}^{T_1} e^{-it^2} dt \quad (8.7)$$

$$= \frac{2iM_1}{\beta - 1} [e^{ih \sin \varphi} - 1] + O(h^3/r^{3/2})$$

where M_1 is obtained by putting $\theta = \pi/2$ in (7.4).

When we combine (8.7) and the expression (7.19) for S_{21} the $2iM_1/(\beta - 1)$ terms cancel leaving

$$(e^{-ix} + S_1)_r + S_{21} = (e^{-ix} + S_1)_\rho + \frac{2iM_1}{\beta - 1} e^{ih \sin \varphi} + i \frac{M_2}{g} \quad (8.8)$$

$$+ O(h^3/r^{3/2}) + O[M_1 \exp(-2\pi h)].$$

The sum of the terms involving M_1 and M_2 may be expressed in a form which contains the expression $c(r)$ defined by (2.5) and the quantity b defined by

$$\begin{aligned}
 b &= -\log \beta = \log \tan \left(\frac{\varphi}{2} + \frac{\pi}{4} \right) = \varphi + \varphi^3/6 + \dots \\
 &= h^{-1/3} g,
 \end{aligned} \tag{8.9}$$

$$\tanh b = \sin \varphi.$$

Replacing $c(r) \exp (ih \sin \varphi)$ by $c(\rho)$ plus a correction term then converts (8.8) into

$$\begin{aligned}
 (e^{-ix} + S_1)_r + S_{21} &= (e^{-ix} + S_1)_\rho \\
 &+ \frac{c(\rho)[1 + \exp (2b_\rho)]^{1/2}}{2^{1/2}} \left[\frac{1}{1 - e^b} + \frac{\exp (ihb - ih \tanh b)}{b} \right]_\rho \\
 &+ 0(h^3/r^{3/2}) + 0(r^{-1/2}e^{-2\pi h}),
 \end{aligned} \tag{8.10}$$

where the subscript ρ on the square brackets indicates that b is to be replaced by b_ρ defined by

$$b_\rho = \log \tan (\psi/2 + \pi/4) = \psi + \psi^3/6 + \dots \tag{8.11}$$

The quantity within the square brackets in (8.10) is continuous at $b = 0$ where it behaves like (neglecting $0(b)$ terms but retaining $0(hb^2)$)

$$\frac{1}{2} + ihb^2/3 = \frac{1}{2} + ih^{1/3}g^2/3. \tag{8.12}$$

Expression (8.10) is to be used with $(S_{22} + S_{23})$ and $(S_{32} + S_{33})$ obtained from (7.17) and (7.46) (with $\theta = \pi/2$ and $w = 1$), respectively.

When ψ is small, expression (8.10) becomes

$$\begin{aligned}
 (e^{-ix} + S_1)_r + S_{21} &= (e^{-ix} + S_1)_\rho \\
 &+ c(\rho) \left[\frac{1}{2} - \frac{1}{\psi} + \frac{h^{1/3} \exp (ig^3/3)}{g} \right]_\rho + \dots
 \end{aligned} \tag{8.13}$$

The subscript ρ on the square bracket indicates that g is to be replaced by g_ρ defined by

$$g_\rho = h^{1/3}b_\rho = h^{1/3}(\psi + \psi^3/6 + \dots). \tag{8.13}$$

When, in accordance with (8.1) and (8.2), we add to (8.13) the approximations (7.26) and (7.49), namely

$$\begin{aligned}
 S_{22} + S_{23} &\sim iM_2\Psi(g), \\
 S_{32} + S_{33} &\sim iM_2\Psi_r(g),
 \end{aligned} \tag{8.14}$$

we obtain

$$E = (e^{-ix} + S_1)_\rho + c(\rho) \left[\frac{1}{2} - \frac{1}{\psi} + \left\{ \frac{1}{g} + \Psi(g) \right\} h^{1/3} \exp(i g^3/3) \right]_\rho \quad (8.15)$$

$$+ \dots,$$

$$H = (e^{-ix} + S_1)_\rho + c(\rho) \left[\frac{1}{2} - \frac{1}{\psi} \right. \\ \left. + \left\{ \frac{1}{g} + \Psi_v(g) \right\} h^{1/3} \exp(i g^3/3) \right]_\rho + \dots \quad (8.16)$$

The terms neglected in (8.15) and (8.16) are the "order of" terms in (8.10), plus those neglected by virtue of ψ being small, plus the errors in (8.14). The errors in (8.14) are of two kinds namely those of $O(h^3/r^{3/2})$ and those due to approximating the parabolic cylinder functions by Airy integrals.

It is interesting to observe the forms assumed by (8.15) and (8.16) when $h = 0$ even though they are not supposed to hold for small values of h . In this case ρ, ψ go into r, φ and the right hand sides of (8.15) and (8.16) become the same, namely

$$(e^{-ix} + S_1)_r + c(r)/2. \quad (8.17)$$

The half plane results given in Section 2 become, for small values of φ ,

$$\frac{E}{H} \Bigg\} = (e^{-ix} + S_1)_r \pm c(r)/2 \quad (8.18)$$

where the upper sign corresponds to E and the lower one to H . Comparison of (8.17) and (8.18) shows that (8.15) for E reduces to the proper value but (8.16) for H fails to do so because the signs of $c(r)/2$ do not agree.

The discrepancy is apparently related to the approximations we have made in obtaining the expression (7.19) for S_{21} from (7.18) and to the errors introduced by approximating the parabolic cylinder functions by the Airy integrals. As we let $h \rightarrow 0$ in the more complete expression (7.18) for S_{21} , the value obtained for $S_{21} \rightarrow \infty$. This is explained by the fact that the upper limit of integration $-1 - ih$ (at point C) approaches the pole of the integrand of (7.17) at $n = -1$. This large value of S_{21} tends to be cancelled by the large value of S_{23} (for horizontal polarization). On the other hand our approximation (7.19) yields via (7.21) the value iM_1 for S_{21} and S_{31} when $h = 0$ and $\beta = 1$. The factor $h^{1/3}$ in M_2 makes our approximations for $S_{22}, S_{23}, S_{32}, S_{33}$ in terms of Airy integrals vanish when $h = 0$.

Incidentally, if, instead of taking the point C of Fig. 6.2 to be at $-1 - ih$, we take it to be at $-\frac{1}{2} - ih$ (a choice which receives some support from the Airy integral representation obtained from the viewpoint of the differential equations discussed in the first part of Section 13), the approximate integrals of (7.17) and (7.46) may be integrated directly when $h = 0$ and $\beta = 1$. It is found that

$$\begin{aligned} S_{21} &\sim -\frac{M_1}{\pi} \log 2, & S_{22} &\sim \frac{M_1}{\pi} \log 2 + \frac{iM_1}{2}, & S_{23} &\sim +iM_1/2, \\ S_{31} &\sim -\frac{M_1}{\pi} \log 2, & S_{32} &\sim \frac{M_1}{\pi} \log 2 - \frac{iM_1}{2}, & S_{33} &\sim -iM_1/2, \end{aligned}$$

and these add to give the values $S_2(0) \sim iM_1$, $S_3(0) \sim -iM_1$ required by the half-plane case.

It is seen that a rather thorough investigation of the errors introduced by our approximations would be required to resolve the discrepancy between (8.17) and (8.18). Since we do not intend to go into this subject, and since the errors we have made may be as large as the $\frac{1}{2}$ which appears within the square brackets of (8.15) and (8.16), we shall "split the difference" between the two polarizations and omit the $\frac{1}{2}$ altogether. This is done in Section 2 where $\tau \approx g_\rho$.

9. THE FUNCTIONS $U_n(z)$, $V_n(z)$, $W_n(z)$

The functions $U_n(z)$, etc., are defined for all values of z and n by the integrals

$$\begin{aligned} U_n(z) &= \frac{1}{2\pi i} \int_U t^{-n-1} e^{-t^2 + 2zt} dt, \\ V_n(z) &= \frac{1}{2\pi i} \int_V t^{-n-1} e^{-t^2 + 2zt} dt, \\ W_n(z) &= \frac{1}{2\pi i} \int_W t^{-n-1} e^{-t^2 + 2zt} dt, \end{aligned} \tag{9.1}$$

where the paths of integration U , V , W in the complex t -plane are shown in Fig. 9.1. The cut in the t -plane runs from $-\infty$ to 0 and has been introduced in order to make the function t^{-n-1} one-valued. In some of the later work the paths of integration will cross this cut. Of course, this requires close attention to $\arg t$.

The initial and final points of the various paths (denoted in Fig. 9.1 by the subscripts i and f) are located at infinity. $\arg t = -\pi$ at U_i and W_f and $+\pi$ at U_f and V_i .

We shall give a summary of the properties of the functions (9.1)

which will be needed in our work. These functions are related to the parabolic cylinder function $D_n(z)$ ^{18,19} through the equations

$$\begin{aligned} U_n(z) &= 2^{n/2} e^{z^2/2} D_n(2^{1/2}z)/\Gamma(n+1), \\ V_n(z) &= -i^{-n} 2^{n/2} e^{z^2/2} D_{-n-1}(-iz2^{1/2})/(2\pi)^{1/2}, \\ W_n(z) &= -i^n 2^{n/2} e^{z^2/2} D_{-n-1}(iz2^{1/2})/(2\pi)^{1/2}. \end{aligned} \quad (9.2)$$

We use the functions $U_n(z)$, etc., here instead of $D_n(z)$ because they seem to be more convenient for the particular problem we have to deal with.

From the definitions (9.1) it follows that $U_n(z)$, $V_n(z)$, $W_n(z)$ are one-valued analytic functions of z and n . By expanding $\exp(2zt)$ in (9.1) and integrating termwise it may be shown that

$$\begin{aligned} U_n(z) &= 2A \cos(\pi n/2) + 4zB \sin(\pi n/2), \\ V_n(z) &= -Ai^{-n} - 2zi^{-n+1}B, \\ W_n(z) &= -Ai^n - 2zi^{n-1}B, \end{aligned} \quad (9.3)$$

$$\begin{aligned} A &= {}_1F_1\left(-\frac{n}{2}; \frac{1}{2}; z^2\right) / 2\Gamma\left(1 + \frac{n}{2}\right), \\ B &= {}_1F_1\left(\frac{1-n}{2}; \frac{3}{2}; z^2\right) / 2\Gamma\left(\frac{1+n}{2}\right). \end{aligned}$$

When $z = 0$ and $U'_n(z) = dU_n(z)/dz$, etc.,

$$\begin{aligned} U_n(0) &= \frac{\cos(\pi n/2)}{\Gamma(1+n/2)}, & V_n(0) &= \frac{-i^{-n}}{2\Gamma(1+n/2)}, \\ W_n(0) &= \frac{-i^n}{2\Gamma(1+n/2)}, \\ U'_n(0) &= \frac{2 \sin(\pi n/2)}{\Gamma\left(\frac{n+1}{2}\right)}, & V'_n(0) &= \frac{-i^{-n+1}}{\Gamma\left(\frac{n+1}{2}\right)}, \\ W'_n(0) &= \frac{-i^{n-1}}{\Gamma\left(\frac{n+1}{2}\right)}. \end{aligned} \quad (9.4)$$

¹⁸ See E. T. Whittaker and G. N. Watson, *Modern Analysis*, Fourth Edition (1927) Cambridge Univ. Press pp. 347-354.

¹⁹ W. Magnus and F. Oberhettinger, *Formeln und Sätze für Speziellen Funktionen*, 2nd Ed., Springer, 1948 Chap. 6 Section 3, and p. 227. A comprehensive account of $D_n(z)$ is given in the forthcoming work, *Higher Transcendental Functions*, compiled by the staff of the Bateman Manuscript Project.

Let $T_n(z)$ denote any one of $U_n(z)$, $V_n(z)$, $W_n(z)$, let primes denote differentiation with respect to z , and let asterisks denote complex conjugates. Then we have the following relations

$$T_n''(z) - 2zT_n'(z) + 2nT_n(z) = 0, \quad (9.5)$$

$$\frac{d}{dz} [e^{-z^2} T_n'(z)] + 2ne^{-z^2} T_n(z) = 0, \quad (9.6)$$

$$\frac{d^2}{dz^2} [e^{-z^2/2} T_n(z)] + (2n + 1 - z^2)e^{-z^2/2} T_n(z) = 0, \quad (9.7)$$

$$T_n'(z) = 2T_{n-1}(z), \quad (9.8)$$

$$nT_n(z) = 2zT_{n-1}(z) - 2T_{n-2}(z),$$

$$\begin{aligned} U_n'(z) V_n(z) - U_n(z) V_n'(z) &= i2^n e^{z^2} / \pi^{1/2} \Gamma(n+1), \\ V_n'(z) W_n(z) - V_n(z) W_n'(z) &= i2^n e^{z^2} / \pi^{1/2} \Gamma(n+1), \\ W_n'(z) U_n(z) - W_n(z) U_n'(z) &= i2^n e^{z^2} / \pi^{1/2} \Gamma(n+1), \end{aligned} \quad (9.9)$$

$$U_n(z) + V_n(z) + W_n(z) = 0, \quad (9.10)$$

$$[V_n(z)]^* = W_n^*(z^*), [W_n(z)]^* = V_n^*(z^*), \quad (9.11)$$

$$[U_n(z)]^* = U_n^*(z^*),$$

$$V_n(-z) = i^{-2n} W_n(z), W_n(-z) = i^{2n} V_n(z), \quad (9.12)$$

$$U_n(-z) = -i^{2n} V_n(z) - i^{-2n} W_n(z),$$

$$V_{-n-1}(iz) = -i^{n+1} K U_n(z), \quad W_{-n-1}(iz) = -i^{-n-1} K U_n(-z),$$

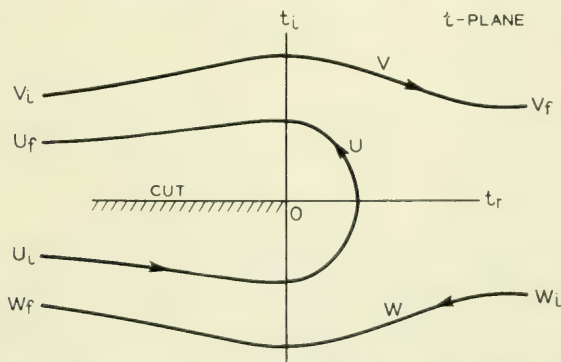


Fig. 9.1 — Paths of integration used in the integrals of equations (9.1) which define the functions $U_n(z)$, $V_n(z)$ and $W_n(z)$. The subscripts i and f stand for the “initial” and “final” points of the paths.

$$\begin{aligned}
 U_{-n-1}(iz) &= i^{-n-1}(i^{2n} - i^{-2n})KW_n(z) = -(2i)^{-n}\pi^{1/2}e^{-z^2}W_n(z)/\Gamma(-n), \\
 U_{-n-1}(-iz) &= i^{n-1}(i^{2n} - i^{-2n})KV_n(z), \\
 K &= e^{-z^2}\Gamma(n+1)/\pi^{1/2}2^{n+1}
 \end{aligned}
 \tag{9.13}$$

Equation (9.5) may be obtained from (9.1) by forming $T_n''(z) - 2zT_n'(z)$ and integrating by parts. Equation (9.10) follows from (9.1) upon joining the paths U , V , W to obtain a closed path of integration. From (9.11) it follows that when we have an expression for $V_n(z)$ which holds for all values of z and n , replacing i by $-i$ (or i^{-1}) gives the corresponding expression for $W_n(z)$. Equations (9.13) may be obtained by using the fact that $U_{-n-1}(iz) \exp(z^2)$ etc. are solutions of the differential equations (9.5) and (9.7).

The relations (9.11) and (9.13) enable us to compute the values of $U_n(z)$, $V_n(z)$, $W_n(z)$ for $z = i^{1/2}\rho$ and $z = i^{-1/2}\rho$ and all n when the values of any two are given for $z = i^{1/2}\rho$ and $\text{Re}(n) \geq -1/2$.

It may be verified that as z becomes large and n remains fixed the differential equation (9.5) has the asymptotic solutions

$$\begin{aligned}
 s_1(n, z) &= \frac{2^n z^n}{\Gamma(n+1)} {}_2F_0\left(-\frac{n}{2}, \frac{1-n}{2}; -1/z^2\right), \\
 s_2(n, z) &= \frac{z^{-n-1}e^{z^2}}{2\sqrt{\pi}} {}_2F_0\left(\frac{n+1}{2}, \frac{n+2}{2}; 1/z^2\right),
 \end{aligned}
 \tag{9.14}$$

where

$$|\arg z| < \pi \text{ and } s_2(-n-1, iz) = i^n K s_1(n, z),$$

K being given by (9.13). In terms of these functions we have

$$\begin{aligned}
 V_n(z) &\sim -is_2(n, z), & 0 < \arg z < \pi \\
 &\sim -s_1(n, z) - is_2(n, z), & -\pi/2 < \arg z < 0 \\
 &\sim -s_1(n, z) - i^{-4n+1}s_2(n, z), & -\pi < \arg z < -\pi/2
 \end{aligned}
 \tag{9.15}$$

$$W_n(z) \sim is_2(n, z), \quad -\pi < \arg z < 0 \tag{9.16}$$

$$U_n(z) \sim s_1(n, z), \quad -\pi/2 < \arg z < \pi/2. \tag{9.17}$$

The first expression for $V_n(z)$ in (9.15) follows when we note that the leading term may be obtained from (9.1) by choosing the path of integration V to be $t = z + \tau$ where τ runs from $-\infty$ to $+\infty$, and $|z|$ is supposed to be large. (9.17) follows from the first of (9.15) and the relation (9.13) between $V_{-n-1}(iz)$ and $U_n(z)$. Asymptotic expressions for $W_n(z)$ may be obtained by taking the conjugate complex of those

for $V_n(z)$. The second and third expressions for $V_n(z)$ follow from the other asymptotic expressions and (9.11), (9.12).

When $R(n) < 0$, the theory of gamma functions and (9.1) lead to

$$\Gamma(n+1)\Gamma(-n)U_n(z) = -\pi \csc \pi n U_n(z) \\ = \int_0^\infty \tau^{-n-1} \exp(-\tau^2 - 2\tau z) d\tau. \quad (9.18)$$

By expressing $\sqrt{\pi} \exp[-(z-t)^2]$ as the integral of

$$\exp[-\tau^2 + 2i(z-t)\tau]$$

taken from $\tau = -\infty$ to $+\infty$ and substituting in (9.1) it may be shown that, when $R(n) > -1$,

$$U_n(z) = F i^n \int_{-\infty}^\infty e^{-t^2 - 2izt} t^n dt, \\ V_n(z) = -F i^{-n} \int_0^\infty e^{-\tau^2 + 2iz\tau} \tau^n d\tau, \quad (9.19) \\ W_n(z) = -F i^n \int_0^\infty e^{-\tau^2 - 2iz\tau} \tau^n d\tau, \\ F = 2^n e^{z^2} / \Gamma(n+1) \pi^{1/2}.$$

When n is not an integer the path of integration in the integral (9.19) for $U_n(z)$ is indented downward at the origin. Equations (9.19) may also be obtained from (9.1) by using (9.13) and (9.18).

When n is an integer

$$U_n(-z) = (-)^n U_n(z), \quad V_n(-z) = (-)^n W_n(z), \quad (9.20)$$

and when n is a positive integer

$$U_n(z) = s_1(n, z) = (-)^n \frac{e^{z^2}}{n!} \frac{d^n}{dz^n} e^{-z^2}, \\ U_{-n}(z) = 0, \quad (9.21)$$

$$V_{-n}(z) = -W_{-n}(z) = -i s_2(-n, z) = -\frac{i}{\sqrt{\pi} 2^n} \frac{d^{n-1}}{dz^{n-1}} e^{z^2}.$$

From Maclaurin's expansion and (9.21),

$$\sum_{n=0}^\infty t^n U_n(z) = \exp[-t^2 + 2zt]. \quad (9.22)$$

10. FORMULAS FOR THE SADDLE-POINT METHOD

Much of our work involves the behavior of the parabolic cylinder functions as functions of n when n is a large complex number. Although this subject has been studied by several writers,^{20, 21, 22} their results are not in the form we require. As the work of Sections 6 and 7 shows, the paths of steepest descent for the integrals in our electromagnetic problem are intimately connected with the function $f(t_0) - f(t_1)$. In turn, this function is closely related to the saddle point method of evaluating $U_n(z)$, etc., for large values of n . For the sake of completeness, we shall outline this method. We shall pay special attention to the relative importance of the two saddle points as n moves about in its complex plane.

When we write the integrand of the integrals (9.1) as $\exp [f(t)]$ we obtain expressions of the form

$$U_n(z) = \frac{1}{2\pi i} \int_U \exp [f(t)] dt, \quad (10.1)$$

$$f(t) = -t^2 + 2zt - m \log t, \quad m = n + 1.$$

The saddle points of the integrand are at the points t_0 and t_1 in the complex t -plane where $f'(t)$ is zero:

$$\begin{aligned} 2t_0^2 - 2zt_0 + m &= 0, & t_0^2 - zt_0 &= -m/2, \\ t_0 &= \frac{z + (z^2 - 2m)^{1/2}}{2}, & t_0 + t_1 &= z, \\ t_1 &= \frac{z - (z^2 - 2m)^{1/2}}{2}, & 2t_0 t_1 &= m. \end{aligned} \quad (10.2)$$

Let the path of integration U of (10.1), for example, be deformed so as to pass through a saddle point, say t_0 , along a path of steepest descent. Let

$$f(t) = f(t_0) - \sum_2^{\infty} b_k(t - t_0)^k/k!. \quad (10.3)$$

Then, if b_2 is not too small, the contribution of the region around t_0

²⁰ Nathan Schwid, The Asymptotic Forms of the Hermite and Weber Functions, Amer. Math. Soc. Trans. **37**, pp. 339-362, 1935. References to earlier work will be found in this paper. Schwid's work is based on R. Langer's study of the asymptotic solutions of second order differential equations.

²¹ O. E. H. Rydbeck, The Propagation of Radio Waves, Trans. of Chalmers Univ. of Tech. **34**, 1944.

²² G. N. Watson, Harmonic Functions Associated with Parabolic Cylinder Functions, Proc. London Math. Soc. (2) **17**, pp. 116-148, 1918.

to the value of the integral is $\exp [f(t_0)]$ times

$$\begin{aligned} & \frac{1}{2\pi} \int \exp \left[- \sum_2^{\infty} b_k (t - t_0)^k / k! \right] dt \\ & \sim (2\pi b_2)^{-1/2} \{ 1 + \frac{1}{2} (-b_4 B_2 + 10 b_3^2 B_3) \\ & + \frac{1}{6} (-b_6 B_3 + [35 b_4^2 + 56 b_3 b_5] B_4 - 2100 b_3^2 b_4 B_5 + 55(280) b_3^4 B_6) \\ & + \dots \} \end{aligned} \quad (10.4)$$

where $B_k = (2b_2)^{-k} / k!$. The sign of $(2\pi b_2)^{-1/2}$ is chosen so that the argument of the right hand side of (10.4) is equal to $\arg (dt)$ at $t = t_0$ on the path of steepest descent. The derivatives of $f(t)$ at t_0 give

$$\begin{aligned} b_2 &= 2(t_0 - t_1)/t_0, & b_3 &= 4t_1/t_0^2, & b_4 &= -12t_1^2/t_0^3, \\ -b_4 B_2 + 10b_3^2 B_3 &= \frac{t_1(t_1 + 9t_0)}{24t_0(t_0 - t_1)^3}. \end{aligned} \quad (10.5)$$

The values of these quantities at the saddle point t_1 may be obtained by interchanging t_0 and t_1 . If more terms of (10.4) are desired they may be obtained from the formal result

$$\begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left[- \sum_{k=2}^{\infty} \alpha_k t^k / k! \right] dt \\ & \sim (2\pi\alpha_2)^{-1/2} \left[1 + \sum_{k=2}^{\infty} Y_{2k}(0, 0, -\alpha_3, -\alpha_4, \dots, -\alpha_{2k}) / k! (2\alpha_2)^k \right] \end{aligned} \quad (10.6)$$

where $Y_n(a_1, a_2, \dots, a_n)$ is the Bell exponential polynomial.²³ It is necessary to rearrange the terms given by (10.6) in order to get them in groups having the same order of magnitude. A more careful treatment of the terms in the asymptotic expansions for $D_n(z)$ has been given by Watson.²² His method is similar to that used by Debye for Bessel functions.

In our work we shall deal with two different complex planes, and the reader is cautioned against confusing them. One is the complex t -plane, shown in Fig. 10.1, which contains the paths of integration for integrals such as (10.1). The other is the complex m -plane, shown in Fig. 10.2, which is introduced because we are often more interested in $U_n(z)$, etc., as functions of $m = n + 1$ than as functions of z . In the earlier sections

²³ E. T. Bell, Exponential Polynomials, Ann. of Math. **35**, pp. 258-279, 1934. The polynomials are tabulated up to $n = 8$ by John Riordan, Inversion Formulas in Normal Variable Mapping, Annals of Math. Stat. **20**, pp. 417-425, 1949.

we have spoken of the complex n -plane, but this is essentially the m -plane shifted by unity.

Since we are going to deal with a fixed value of z ($i^{1/2} \xi$ or $i^{-1/2} \eta$) but with a variable value of m , we make t_0 and t_1 one-valued functions of m by cutting the m -plane as shown in Fig. 10.2.

It may be shown that t_0 and t_1 lie in the opposite half-planes indicated in Fig. 10.1. This restricts $\arg t_0$ to lie between $\arg z - \pi/2$ and $\arg z + \pi/2$. $\arg t_1$ is restricted to lie between $\arg z - \pi$ and $\arg z + \pi$ by the cut shown in Fig. 10.1. It may also be shown that

$$|t_0| \geq |t_1|, \quad |\arg t_0 - \arg t_1| \leq \pi. \quad (10.7)$$

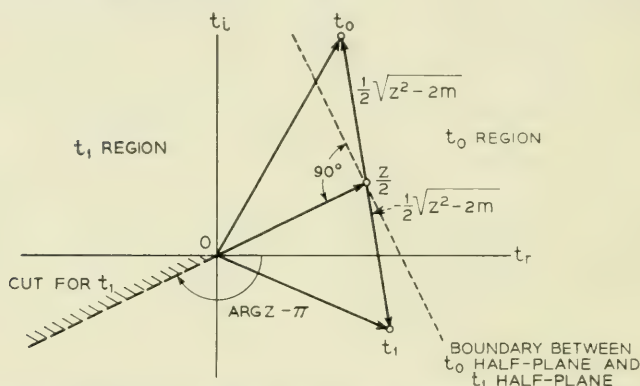


Fig. 10.1 — Diagram showing the half-plane regions to which the saddle points t_0 and t_1 are confined in the t -plane.

One might wonder why cuts in the m -plane are required since it has already been pointed out that $U_n(z)$, etc., are one-valued functions of $m = n + 1$. The trouble is that the asymptotic expressions for $U_n(z)$ are many-valued functions of m even though $U_n(z)$ itself is not.

Now that we have considered the saddle points t_0 and t_1 , we turn to a consideration of the paths of steepest descent in the t -plane which pass through them.* The path of steepest descent which passes through t_0 , for example, is that branch of the curve

$$\text{Im} [f(t) - f(t_0)] = 0 \quad (10.8)$$

for which t_0 is the highest point (i.e., $\text{Re} [f(t) - f(t_0)] \leq 0$ on it). The

* Watson²² has studied paths corresponding to $\text{Re}(n) > 0$ when z is any complex number, and has given curves which are related to some of those shown in Section 11.

paths of steepest descent may be shown to have the following properties:

1. Let $t = t_r + it_i = r \exp(i\theta)$. Then the paths of steepest descent either run out to $t_r = \pm \infty$ with $t_i \rightarrow \text{Im } z$ or spiral in to $t = 0$ as $r = (\text{constant}) \exp(-m_r\theta/m_i)$.

2. The steepest descent path through t_0 may be computed by a graphical method based on*

$$\arg(dt) = \arg t - \arg(t - t_0) - \arg(t - t_1). \quad (10.9)$$

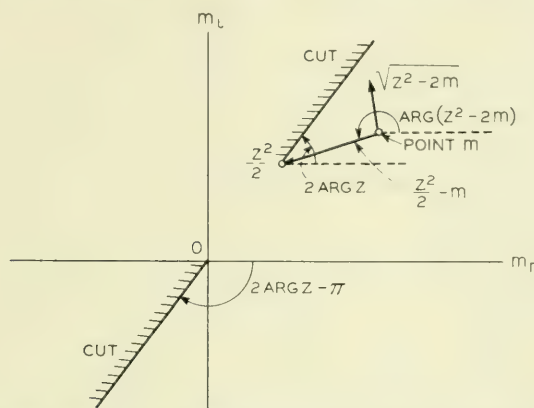


Fig. 10.2 - Diagram showing the cuts in the complex m -plane, $m = n + 1$.

If we draw the triangle $t_0 0 t_1$ and bisect the interior angle at t_0 by the line $b_0 t_0$ then

$$\arg(dt) \text{ at } t_0 = \text{angle } t_1 t_0 b_0. \quad (10.10)$$

If one goes clockwise in traveling from the side $t_0 t_1$ to $t_0 b_0$ then $\arg dt$ is negative. Likewise, $\arg(dt)$ at t_1 (on the path through t_1) is the angle between the side $t_1 t_0$ and the bisector $t_1 b_1$ of the interior angle at t_1 .

3. When m has the critical value $z^2/2$ the saddle points coincide: $t_0 = t_1 = z/2$, and the paths of steepest descent start out from $t = z/2$ in the three directions $\arg(t - z/2) = (\arg z)/3 + \delta$ where δ is 0, $2\pi/3$, or $-2\pi/3$.

4. Some of the paths of steepest descent change their character as m goes from one region of the m -plane to another. This is illustrated in Section 11 for the case $z = \rho \exp(i\pi/4)$ where it is shown that the

* A similar method was used in 1938 by A. Erdélyi in an unpublished study of the asymptotic behavior of confluent hypergeometric functions.

boundaries are given by

$$\operatorname{Im} [f(t_0) - f(t_1)] = 0, \quad (10.11)$$

or a similar equation involving another pair of saddle points, e.g., t_1 and $t_1 \exp(i2\pi)$. In this equation z is regarded as fixed and t_0, t_1 are functions of m defined by (10.2). It should be noted that although (10.8) defines a path of steepest descent in the t -plane, (10.11) defines curves (boundaries of regions) in the m -plane.

5. If m is such that the path of integration for a particular function, say $U_n(z)$, passes through both t_0 and t_1 , each one will contribute to the value of $U_n(z)$. Furthermore, if m is such that

$$\operatorname{Re} [f(t_0) - f(t_1)] = 0, \quad (10.12)$$

t_0 and t_1 have the same height and the two contributions have a chance of cancelling each other and giving a value of zero for $U_n(z)$. Thus (10.12) or some similar equation defines the lines in the m -plane along which the zeros of $U_n(z)$, etc., (regarded as functions of m) are asymptotically distributed.

6. The lines in the m -plane defined by (10.11) and (10.12) may be obtained by substituting the values (10.2) for t_0 and t_1 in

$$f(t_0) - f(t_1) = t_0^2 - t_1^2 - 2t_0t_1 \log(t_0/t_1), \quad (10.13)$$

and setting the imaginary and real parts, respectively, to zero. However, instead of dealing with m directly it is easier to use $w = u + iv$ defined by

$$w = \log(t_0/t_1) = \log|t_0/t_1| + i(\arg t_0 - \arg t_1), \quad (10.14)$$

$$m = z^2/(\cosh w + 1), \quad (10.15)$$

where (10.15) follows from (10.14) and (10.2). Then (10.13) becomes

$$\begin{aligned} f(t_0) - f(t_1) &= m(\sinh w - w) \\ &= \frac{z^2(\sinh w - w)}{\cosh w + 1}. \end{aligned} \quad (10.16)$$

The inequalities (10.7) show that

$$u \geq 0, |v| \leq \pi.$$

7. For the special case $z = \rho \exp(i\pi/4)$, (10.16) gives

$$(\cosh u + \cos v - v \sin v) \sinh u = (\cosh u \cos v + 1) u, \quad (10.17)$$

$$(\cos v + \cosh u + u \sinh u) \sin v = (\cosh u \cos v + 1) v,$$

respectively, for $Im[f(t_0) - f(t_1)] = 0$ and $Re[f(t_0) - f(t_1)] = 0$. These equations are plotted in Fig. 10.3. It will be noted that a curve is shown for $v > \pi$ even though this puts w outside the allowed rectangle. This is done because one of the paths of integration, W , passes through both t_0 and $t_1 \exp(-i2\pi)$ when m is in a certain region, and the corresponding zeros of $W_n(z)$ lie on the curve defined by

$$Re[f(t_0) - f(t_1 \exp\{-i2\pi\})] = 0.$$

It may be shown that a curve corresponding to

$$f(t_0) - f(t_1 \exp\{-i2\pi\})$$

with $-\pi < v < \pi$ may be obtained from the curve corresponding to $f(t_0) - f(t_1)$ with $\pi < v < 3\pi$ by simply subtracting 2π from v . This is done on Fig. 10.3.

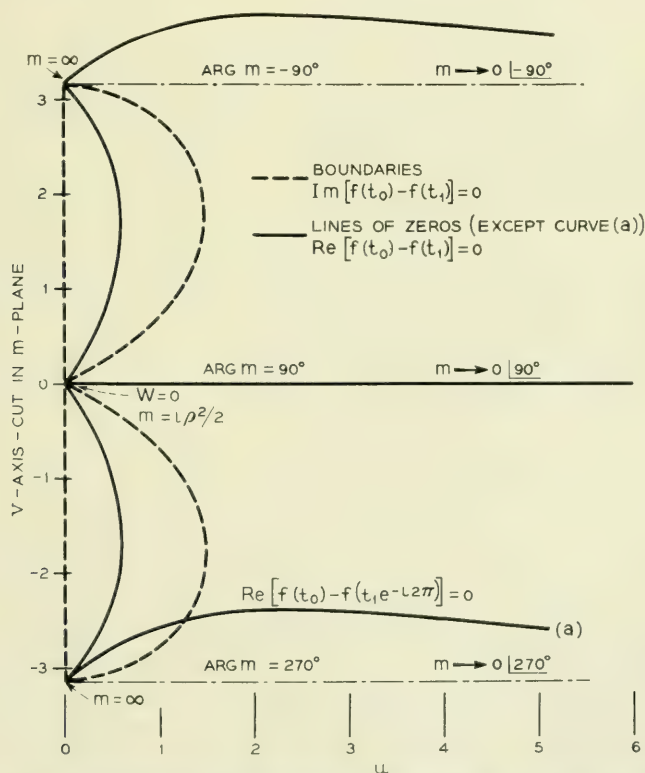


Fig. 10.3 - Boundaries of the regions shown in Fig. 11.2 and lines of zeros shown in Fig. 12.1 as they appear on the $w = u + iv$ plane when $z = i^{1/2}\rho$.

11. DESCRIPTION OF PATHS OF STEEPEST DESCENT

In this section we shall study the paths of steepest descent when $z = i^{1/2}\rho$ in some detail because it is one of the two cases (the other is $i^{-1/2}\rho$) encountered in our diffraction problem. Curves related to some of those shown here have been studied by Watson (for $Re(n) > 0$, as mentioned in Section 10) and by Rydbeck* (for $Re(n) = -1/2$).

The affixes t_0 and t_1 of the saddle points are given by (10.2) and are made definite by the two cuts in the m -plane which now run from $m = i\rho^2/2$ to $m = i\infty$, and from $m = 0$ to $m = -i\infty$. From Figs. 10.1 and 10.2 (drawn for $z = i^{1/2}\rho$) we obtain,

$$\begin{aligned} -\pi/4 &< \arg(i\rho^2 - 2m)^{1/2} \leq 3\pi/4, \\ -\pi/2 &< \arg m \leq 3\pi/2, \\ -\pi/4 &< \arg t_0 \leq 3\pi/4, \\ -3\pi/4 &< \arg t_1 \leq 5\pi/4. \end{aligned} \tag{11.1}$$

A convenient equation for the path of steepest descent through t_0 is obtained by setting $t = r \exp(i\theta)$, $n + 1 = m = m_r + im_i$ in (10.1) and (10.8), and dividing through by ρ^2 :

$$\begin{aligned} - (r/\rho)^2 \sin 2\theta + 2(r/\rho) \sin(\theta + \pi/4) - (m_i/\rho^2) \log(r/\rho) \\ - (m_r/\rho^2)\theta = \text{Im}[f(t_0) + m \log \rho]/\rho^2. \end{aligned} \tag{11.2}$$

Replacing t_0 by t_1 gives the equation for the path through t_1 . Equation (11.2) and its analogue for t_1 were used to compute the paths of steepest descent shown in Figs. 11.1 and 11.6.

When m/ρ^2 is small, so is t_1/ρ . It may be shown from (11.2) (for t_1) that

$$\begin{aligned} (r/|t_1|) \sin(\theta + \pi/4) - (m_i/|m|) \log(r/|t_1|) \\ - \theta m_r/|m| \approx (m_i - m_r \arg t_1)/|m| \end{aligned} \tag{11.3}$$

gives the behavior near $t = 0$ of the path through t_1 . Paths computed from (11.3) are shown in Figs. 11.3 and 11.5. Here $t_1 \approx mi^{-1/2}/2\rho$.

In computing the paths shown by the figures of this section, the work was simplified by taking m to be purely real or purely imaginary, or by assuming m/ρ^2 to be small. Even so, this often required the solution of a rather simple transcendental equation [as (11.2) and (11.3) show]. The graphical method based on (10.9) was not used, although it might

* Pages 26-36 of Reference 21 cited on page 478.

have been if we had elected to study a case in which neither m_r nor m_i were zero.

Most of the values of m/ρ^2 studied in this section are listed in Table 11.1. The point numbers are those listed below and shown in Fig. 11.2. Paths of steepest descent are shown for all but the last two entries of the table. We now take up these values of m one by one.

1. $m = \rho^2/2$. Fig. 11.1 (a). Since $\text{Re } f(t_1) > \text{Re } f(t_0)$, t_1 is higher than t_0 . In all of the figures dealing with the t -plane in this section, solid lines mean $|\arg t| < \pi$ while dashes indicate $|\arg t| > \pi$.

2. $m = \rho^2$. Fig. 11.1 (c). As m goes from $\rho^2/2$ to ρ^2 the path through t_1 changes its type. t_1 is higher than t_0 .

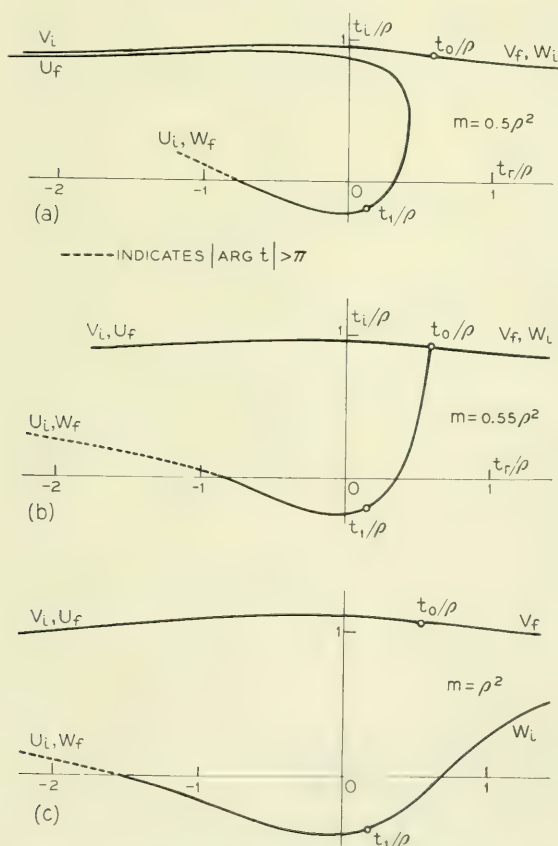


Fig. 11.1 — Paths of steepest descent in $t = t_r + it_i$ plane when $z = i^{1/2}\rho$ and $m \equiv n + 1$ is real and positive. U_i and U_F denote initial and final branches of the path of integration for $U_n(i^{1/2}\rho)$, and so on. t_0 and t_1 are saddle points.

TABLE 11.1 — SADDLE POINT VALUES FOR REPRESENTATIVE VALUES OF m WHEN $z = i^{1/2}\rho$

Point Number	m/ρ^2		t_0/ρ		t_1/ρ		$f(t_0) + m \log \rho / \rho^2$		$f(t_1) + m \log \rho / \rho^2$	
	Mod.	Arg.	Mod.	Arg.	Mod.	Arg.				
1	0.5	0°	1.07	57°	0.23	-57°	-0.013	+ i0.550	1.21	+ i0.450
2	1	0	1.20	64	0.42	-64	-0.080	+ i0.021	1.77	+ i0.979
3	0.550	0	1.09	58	0.25	-58	-0.017	+ i0.500	1.28	+ i0.500
4	0.005	0	1.00	45.1	0.0025	-45.1	-0.000	+ i0.996	0.035	+ i0.004
6	0.005	90	0.997	45	0.0025	45	0.004	+ i1.000	0.004	+ i0.035
8	0.005	180	1.00	44.9	0.0025	135.1	+0.000	+ i1.004	-0.035	+ i0.012
9	0.005	270	1.00	45	0.0025	225	-0.004	+ i1.000	-0.020	+ i0.035
10	0.005	-90	1.00	45	0.0025	-135	-0.004	+ i1.000	0.012	+ i0.035
11	0.550	180	1.09	32	0.25	148	0.017	+ i1.36	-1.28	+ i1.36
12	1	180	1.20	26	0.42	154	0.077	+ i1.59	-1.77	+ i2.55
13	1	90- ϵ	0.71	90	0.71	0	1.07	+ i1.35	0.5	+ i1.35
--	1	$\arg(z^2 - 2m) = 270$	0.71	0	0.71	90	0.5	+ i1.35	1.07	+ i1.35
14	0.5	$\arg(z^2 - 2m) = -90$	0.5	45	0.5	45	0.39	+ i1.10	0.39	+ i1.10
--	1	90	1.37	45	0.37	225	-0.784	+ i1.18	-3.93	+ i1.87
--	1	270	1.37	45	0.37	-135	-0.784	+ i1.18	+2.35	+ i1.87
--	1	-90	1.37	45	0.37	-135	-0.784	+ i1.18	+2.35	+ i1.87

The entries in this table must satisfy $t_0/\rho + t_1/\rho = i^{1/2}$. Equations (12.8) show that when all of the entries are replaced by their complex conjugates, a table for $z = i^{-1/2}\rho$ is obtained.

3. $m = (0.54953 \dots) \rho^2 \sim 0.55 \rho^2$. Fig. 11.1 (b). This value of m marks the change in type of path. Since $\text{Im } f(t_1) = \text{Im } f(t_0)$ is satisfied, the paths through t_0 and t_1 have the same equation [see (11.2)], and there is a chance for a situation like that at t_0 in Fig. 11.1 (b) to occur. The high point of the path U is at t_1 , and it goes continually downhill on either side of t_1 although its direction changes sharply by 90° at t_0 . The point $m = 0.55 \rho^2$ is just one point on the boundary between regions in the m -plane corresponding to various types of paths. The boundary lines are obtained by solving condition (10.11) for m as outlined in Items 6 and 7 of Section 10. Mapping the boundary lines

$$\text{Im } [f(t_0) - f(t_1)] = 0$$

from the auxiliary w -plane (shown in Fig. 10.3) to the m -plane with the help of $m = i\rho^2/(\cosh w + 1)$ gives the boundaries between the regions

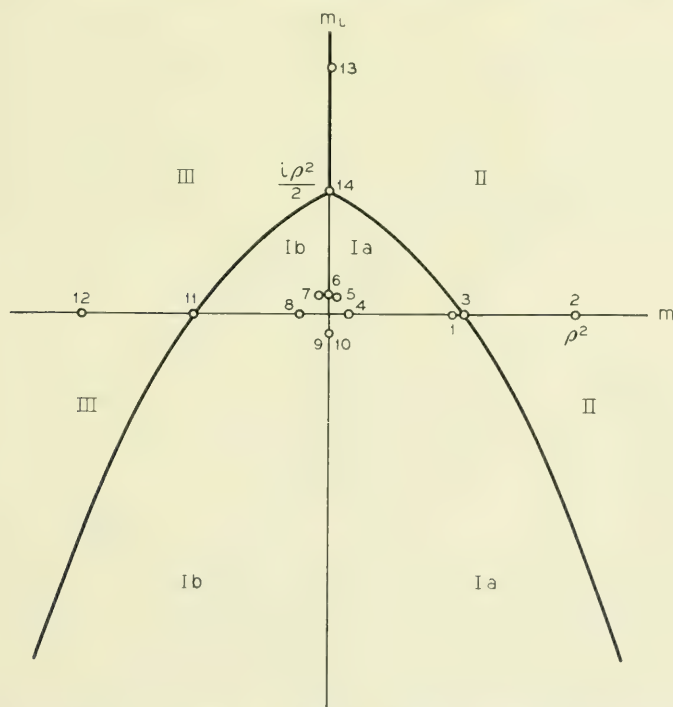


Fig. 11.2 — Regions of different types of paths of steepest descent, and hence different types of asymptotic expansions, when $z = i^{1/2}\rho$. Points numbered 1, 2, 3, are values of m corresponding to the paths of Figs. 11.1 (a), (c), (b). Points designated by 5, 6, 7 correspond to Fig. 11.3(b). Points 4, 8, 9 correspond to Fig. 11.5 and points 11, 12, 13, 14 to Figs. 11.6 (a), (b), (c), (d).

I, II, III shown in Fig. 11.2. It may be verified that for large negative values of m_i the boundary lines in Fig. 11.2 are given approximately by

$$m_r = \pm 2^{3/2} \pi^{-1} \rho |m_i|^{1/2}. \quad (11.4)$$

There was a period, while these curves were being worked out, during which it appeared that the regions I, II, III told the entire story. However, when small values of m were studied it was found that region I splits up into the two sub-regions, Ia and Ib, such that the boundary between them is given by

$$\text{Im} [f(t_1) - f(t_1 e^{-2\pi i})] = 0. \quad (11.5)$$

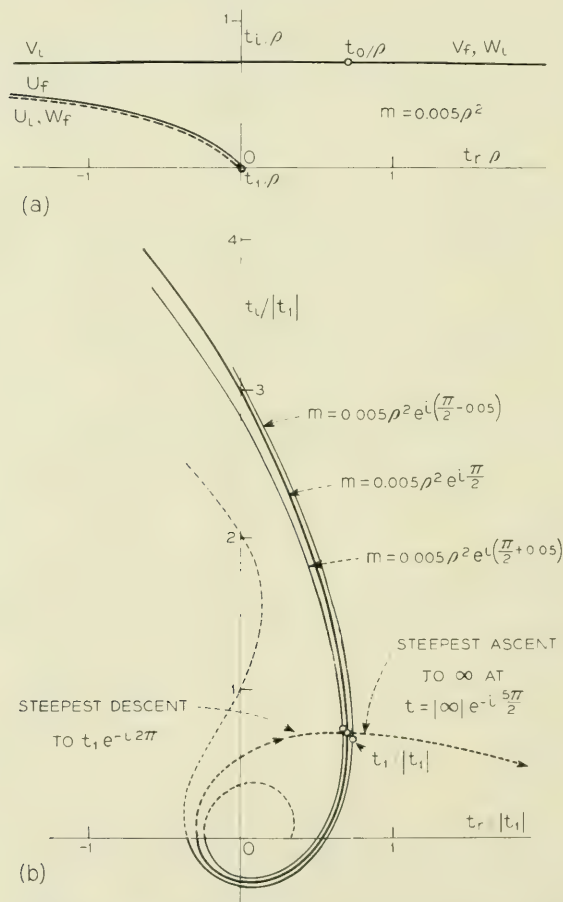


Fig. 11.3 — Paths of steepest descent for $|m| = 0.005 \rho^2$, $z = i^{1/2} \rho$. Away from $t = t_1$ all paths look much like Fig. 11.3(a).

This is of the same form as (10.11). That $t_1 \exp(-2\pi i)$ is a saddle point follows from differentiation of the equation

$$f(te^{-2\pi i}) = f(t) + 2\pi i m_r - 2\pi m_i. \quad (11.6)$$

Combining (11.5) and (11.6) shows that the boundary between Ia and Ib is given by $m_r = 0$. This is indicated on Fig. 11.2.

We now examine the paths of steepest descent when m is small. Fig. 11.3 (a) gives a large view of all the paths, irrespective of $\arg m$, when m^2/ρ is small.

4. $m = 0.005\rho^2$. Fig. 11.5 shows the vicinity around t_1 .

5. $|m| = 0.005\rho^2$, $\arg m = \pi/2 - 0.05$. Fig. 11.3 (b).

6. $|m| = 0.005\rho^2$, $\arg m = \pi/2$. Fig. 11.3 (b). After passing through t_1 the path encircles the origin clockwise and runs down into the saddle point at $t = t_1 \exp(-2\pi i)$. Since m_i is positive, (11.6) shows that $t_1 \exp(-2\pi i)$ is lower than t_1 . The path for $\arg m = \pi/2 - 0.05$ suggests that from $t_1 \exp(-2\pi i)$ the path runs out to $\infty \exp(-\pi i)$ along the path of steepest descent which lies directly under (on the Riemann sheet for $-3\pi < \arg t < -\pi$) the path which runs from t_1 to $t = \infty \exp(i\pi)$. It follows from (11.6) that, as t traces out a path of steepest descent through t_1 , $t \exp(-2\pi i)$ traces out a path of steepest descent through $t_1 \exp(-2\pi i)$ directly under the path through t_1 .

7. $|m| = 0.055\rho^2$, $\arg m = \pi/2 + 0.05$. Fig. 11.3 (b) shows that after passing through t_1 the path of steepest descent spirals in to $t = 0$. According to (11.3), the spiral is given by

$$r \approx (\text{constant}) \exp(-m_r \theta / m_i) \quad (11.7)$$

when r is small and θ large. Two things are to be noted. First, the type of path is different from that for $\arg m = \pi/2 - 0.05$. Hence $\arg m = \pi/2$ marks a change of type similar to that shown in Fig. 11.1 (b), except that here $t_1 \exp(-2\pi i)$ takes the place of t_0 . Condition (11.5) takes the place of condition (10.11), and is satisfied by virtue of $m_r = 0$ when $\arg m = \pi/2$.

The second thing to be noted is that up until now all of the paths of steepest descent have ended at $\pm \infty$ and U , V , W could be deformed into them without difficulty. How can we deform U , for example, into a path of steepest descent when the path through t_1 spirals in to $t = 0$? The way to deal with this problem is shown in Fig. 11.4 where U is continuously deformed into two portions, one coinciding with the path through t_1 , as shown in Fig. 11.3 (b), and the other with the path of steepest descent through $t_1 \exp(-2\pi i)$. The second portion lies directly "underneath" the first portion.

In Fig. 11.4 the dashes mean, as before, that the path of steepest descent is on a sheet of the Riemann surface other than $|\arg t| < \pi$. The alternate dots and dashes are used to indicate that $|\arg t| > \pi$ and that in addition the path lies directly under the curve it parallels. Although in Fig. 11.4 the two kinds of dashed curves are joined at about $\arg t = -3\pi - \pi/4$, they actually should spiral in to $t = 0$ before they connect.

8. $|m| = 0.005\rho^2$, $\arg m = \pi$. Fig. 11.5. For $\arg m = \pi$, (11.6) shows that t_1 and $t_1 \exp(-2\pi i)$ are of the same height.

9. $m = 0.005\rho^2$, $\arg m = 3\pi/2$. For $\pi < \arg m < 3\pi/2$, $t_1 \exp(-2\pi i)$ is higher than t_1 and the paths spiral into $t = 0$ counterclockwise. At $\arg m = 3\pi/2$ the rate of spiralling is zero and we have the path shown in Fig. 11.5 (which is the path for $\arg m = \pi/2$ rotated by 180 degrees). Here $\arg t_1 = 5\pi/4$.

10. $m = 0.005\rho^2$, $\arg m = -\pi/2$. The paths for $\arg m$ equal to $-\pi/2$

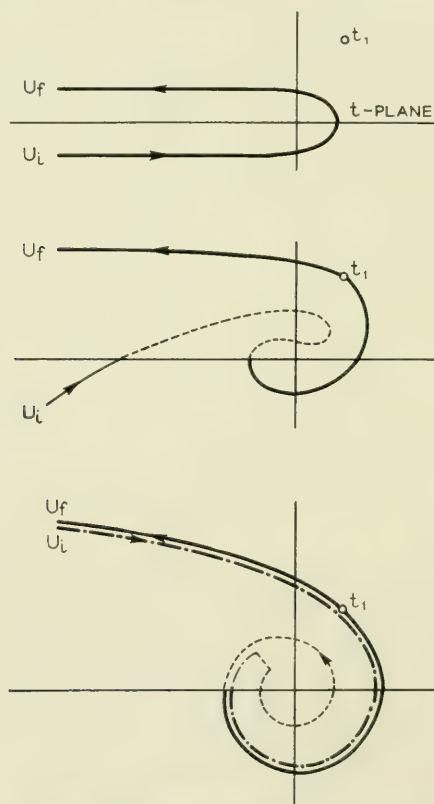


Fig. 11.4 — Deformation of path of integration U into path of steepest descent through t_1 when $m = 0.005 \rho^2 \exp(i\pi/2 + i0.05)$.

and $3\pi/2$ have the same shape and both are highest at the saddle point whose argument is $-3\pi/4$. In both cases the contributions are the same and hence the value of $U_n(z)$, for example, is the same for $\arg m = -\pi/2$ as for $3\pi/2$ (as it must be since our parabolic cylinder functions are one-valued functions of m).

Before leaving the region around $m = 0$ we point out that when $|m/\rho^2| \ll 1$ the path of steepest descent through t_0 is almost independent of $\arg m$. Also, the curves of steepest descent for

$$1/\Gamma(m) = \frac{1}{2\pi i} \int_U e^t t^{-m} dt \quad (11.8)$$

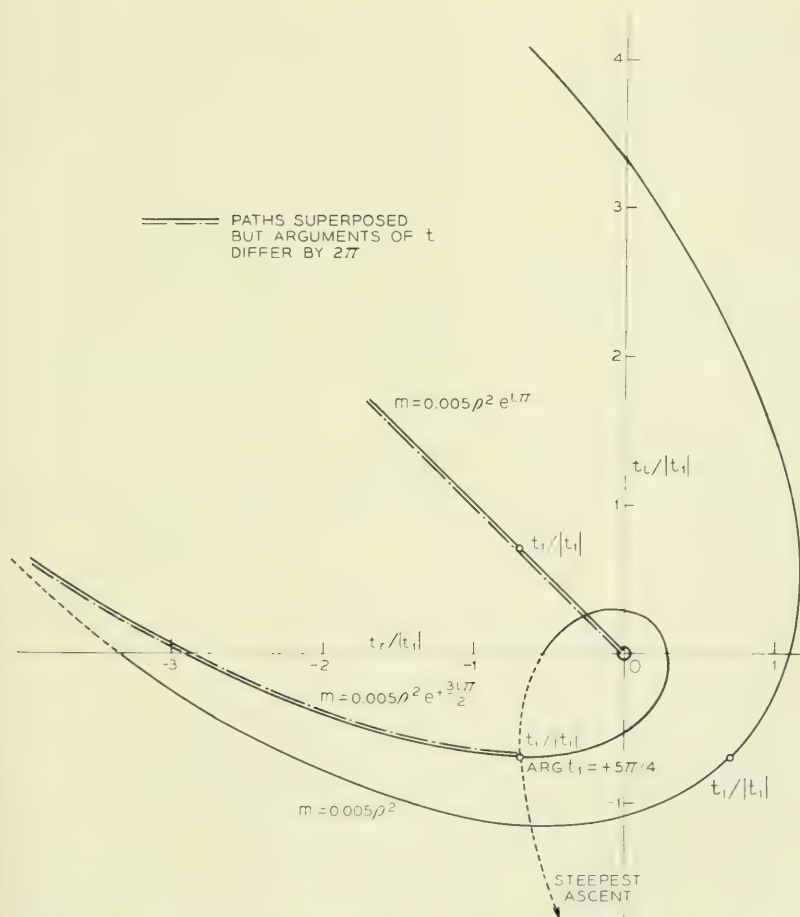


Fig. 11.5 — Paths of steepest descent for $|m| = 0.005 \rho^2$, $z = i^{1/2}\rho$. These curves are much the same as those in Fig. 11.3(b) except for the values of $\arg m$.

behave in much the same way as those just described. The line $m_r = 0$ divides the m -plane into two regions corresponding to different types of paths, and the negative real axis is a line of zeros corresponding to $\text{Re} [f(t_1) - f(t_1 \exp(-2\pi i))] = 0$ where $t_1 = m$ is the saddle point.

11. $|m| = 0.55\rho^2$, $\arg m = \pi$. Fig. 11.6 (a). This value of m marks a change in the type of path.

12. $|m| = \rho^2$, $\arg m = \pi$. Fig. 11.6 (b).

13. $|m| = \rho^2$, $\arg m = \pi/2$, $\arg(i\rho^2 - 2m) = 3\pi/2$. Fig. 11.6 (c). The complication of the paths in Fig. 11.6 (c) is due to the superposition of two boundaries in the m -plane. $\text{Im} f(t_0) = \text{Im} f(t_1)$ accounts for

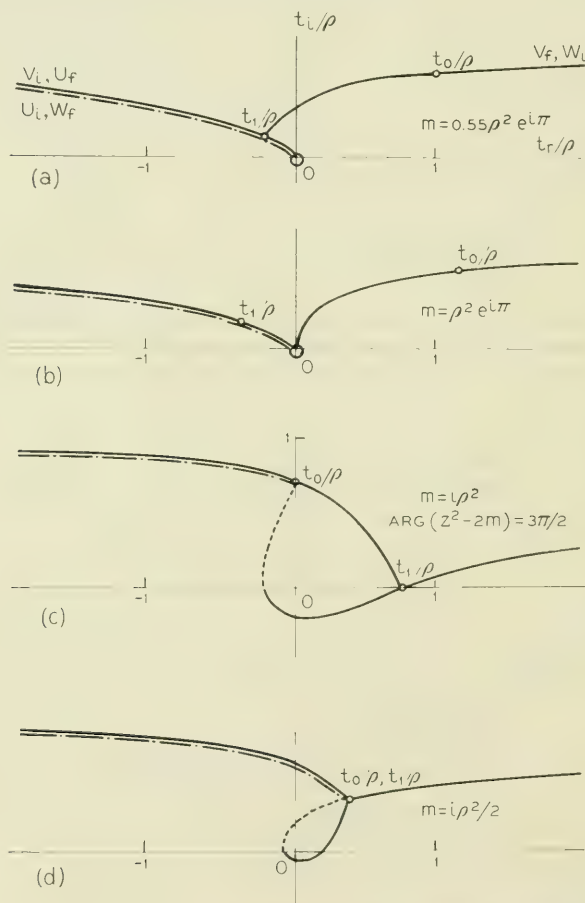


Fig. 11.6 — Paths of steepest descent for miscellaneous values of m with $z = i^{1/2}\rho$.

the path running from t_0 to t_1 , and $\text{Im } f(t_1) = \text{Im } f[t_0 \exp(-2\pi i)]$ for the one running from t_1 to $t_0 \exp(-2\pi i)$. The saddle points in order of their height are t_0 , t_1 , $t_0 \exp(-2\pi i)$, t_0 being the highest.

14. $m = i\rho^2/2$. Fig. 11.6 (d). Here $t_0 = t_1$ and the dashed lines go into the saddle point at $t_1 \exp(-2\pi i)$. The paths of steepest descent change their directions upon passing through the saddle points.

12. ASYMPTOTIC EXPRESSIONS FOR $U_n(z)$, $V_n(z)$, $W_n(z)$

The asymptotic expressions given here are for $z = i^{1/2}\rho$ and $z = i^{-1/2}\rho$ [with $i^{1/2} = \exp(i\pi/4)$] when n is not too close to $z^2/2$. As mentioned earlier, there is a close relation between our results and those given by Schwid.* The main difference is that we regard n as variable and z as fixed while Schwid regards z as variable with n fixed. Another point of difference is that in place of the $m = n + 1$ which appears in our expressions for t_0 and t_1 the quantity $n + 1/2$ appears in Schwid's work. The quantity $2n + 1$ appears to enter naturally when the asymptotic values are obtained from the differential equations. This is seen when the WKB method is applied to equation (9.7).

By examining the paths of steepest descent shown in the figures of Section 11 we can determine the saddle points corresponding to $U_n(z)$, etc., (for $z = i^{1/2}\rho$) for various values of n . The contributions to the integral (10.1) from the saddle points t_0 to t_1 were discussed in Section 10. The contribution from the saddle point $t_1 \exp(-2\pi i)$ (which enters when $z = i^{1/2}\rho$) is, from (11.6), $\exp(i2\pi m)$ times the contribution from t_1 .

Although we shall be concerned mainly with asymptotic expressions for the parabolic cylinder functions themselves, expressions for their derivatives may be readily obtained. Thus $U'_n(z) = dU_n(z)/dz$ has the asymptotic expression

$$\begin{aligned} U'_n(z) &\sim 2t_0 [\text{contribution of } t_0 \text{ to } U_n(z)] \\ &\quad + 2t_1 [\text{contribution of } t_1 \text{ to } U_n(z)] \\ &\quad + 2t_1 [\text{contribution of } t_1 \exp(-2\pi i) \text{ to } U_n(z)] \end{aligned} \quad (12.1)$$

and similar expressions hold for $V'_n(z)$, $W'_n(z)$. These follow when we note that differentiation of the integrals (9.1), which define the functions, introduces a factor $2t$ into the integrand. Of course, if the path of integration does not pass through a particular saddle point, its contribution to (12.1) is zero. Upon replacing t_0 and t_1 by their expressions

* Reference 20, page 478.

(10.2) and subtracting the corresponding expression for $zU_n(z)$ we obtain

$$\begin{aligned} {}'U_n(z) &= U_n'(z) - z U_n(z) \\ &\sim (z^2 - 2m)^{1/2} [(t_0 \text{ contribution}) - (t_1 \text{ contribution}) \\ &\quad - (t_1 e^{-2\pi i} \text{ contribution})] \end{aligned} \quad (12.2)$$

where $'U_n(z)$ is the function defined by (4.19). The same is true for $'V_n(z)$ and $'W_n(z)$.

Consideration of the various paths of integration shown in Section 11 leads to the results shown in Table 12.1. The leading terms of the asymptotic expansions are listed for the various regions of the m -plane

TABLE 12.1 — LEADING TERMS IN THE ASYMPTOTIC EXPANSIONS FOR $U_n(z)$, $V_n(z)$, $W_n(z)$ WHEN $z = i^{1/2}\rho$, $\rho > 0$

Region in m -plane $m = n + 1$	$U_n(i^{1/2}\rho)$	$V_n(i^{1/2}\rho)$	$W_n(i^{1/2}\rho)$
Ia	A_1	A_0	$-A_0 - A_1$
II	$A_1 - A_0$	A_0	$-A_1$
Ib	$(1 - i^{4n})A_1$	A_0	$-A_0 - A_1 + i^{4n}A_1$
III	$(1 - i^{4n})A_1$	$A_0 - A_1$	$-A_0 + i^{4n}A_1$

shown in Fig. 11.2. If the next order terms are required, they may be obtained from (10.4) and (10.5).

The notation used in Table 12.1 is as follows:

$$\begin{aligned} z &= i^{1/2}\rho, \quad m = n + 1, \quad i = \exp(i\pi/2), \\ -\pi/2 &< \arg m \leq 3\pi/2, \quad -\pi/4 < \arg t_0 \leq 3\pi/4, \\ -\pi/2 &< \arg(i\rho^2 - 2m) \leq 3\pi/2, \quad -3\pi/4 < \arg t_1 \leq 5\pi/4, \\ t_0 &= [i^{1/2}\rho + (i\rho^2 - 2m)^{1/2}]/2, \quad t_1 = [i^{1/2}\rho - (i\rho^2 - 2m)^{1/2}]/2, \\ A_0 &= [t_0^{1/2}(i\rho^2 - 2m)^{-1/4}/2i\pi^{1/2}] \exp f(t_0), \\ A_1 &= [t_1^{1/2}(i\rho^2 - 2m)^{-1/4}/2\pi^{1/2}] \exp f(t_1), \\ f(t_0) &= zt_0 + \frac{m}{2} - m \log t_0 = \frac{m}{2} \left(1 - \log \frac{m}{2} - \log \frac{t_0}{t_1} \right) + i^{1/2}\rho t_0, \\ f(t_1) &= zt_1 + \frac{m}{2} - m \log t_1 = \frac{m}{2} \left(1 - \log \frac{m}{2} - \log \frac{t_1}{t_0} \right) + i^{1/2}\rho t_1. \end{aligned} \quad (12.3)$$

Sometimes it is helpful to use

$$\begin{aligned}
 (2\pi)^{-1/2} \exp \left[\frac{m}{2} \left(1 - \log \frac{m}{2} \right) \right] \\
 \sim \begin{cases} 1/\Gamma \left(\frac{m+1}{2} \right) = 1/\Gamma(1+n/2) & \text{for} \\ -\pi/2 < \arg m < \pi/2, & (12.4) \\ i^{-n} \Gamma \left(\frac{1-m}{2} \right) / 2\pi = i^{-n-1} \Gamma(-n/2) / 2\pi, \\ \pi/2 < \arg m < 3\pi/2, \end{cases}
 \end{aligned}$$

where the last line is obtained by setting $m \exp(-\pi i)$ for m in the second line.

The asymptotic expansions for regions *Ib* and *III* may be obtained from those for *Ia* and *II* by using equations (9.11) and (9.13). However, the work is more difficult than one might suspect at first glance.

Incidentally, the leading terms in the asymptotic expansions (9.15) and (9.17), which hold when $\rho \rightarrow \infty$ and n remains fixed, may be obtained by considering the entries for *Ia* and *Ib* in Table 12.1.

It is sometimes convenient to use the limiting forms of the asymptotic expressions when $|m| \gg \rho^2$. In this case, for $z = i^{1/2}\rho$,

$$\begin{aligned}
 2'_{0'} &\rightarrow z \pm i(2m)^{1/2} \left[1 - \frac{z^2}{2m} \right] + O(m^{-3/2}), \\
 2'_{1'} &\rightarrow z \mp i(2m)^{1/2} \left[1 - \frac{z^2}{2m} \right] + O(m^{-3/2}), \\
 \log t_1/t_0 &\rightarrow \mp i\pi \pm iz(2m)^{-1/2} \left[2 + \frac{z^2}{6m} \right] + O(m^{-5/2}),
 \end{aligned} \tag{12.5}$$

where the upper signs hold when $-\pi/2 < \arg m < \pi/2$ and the lower ones when $\pi/2 < \arg m < 3\pi/2$. Substituting (12.5) in (12.3), neglecting the higher order terms, and setting

$$\begin{aligned}
 B &= 2^{-3/2} \pi^{-1/2} \exp \left[\frac{m}{2} \left(1 - \log \frac{m}{2} \right) + i\rho^2/2 \right], \\
 \alpha_0 &= \exp [-\rho(2m/i)^{1/2}], \\
 \alpha_1 &= \exp [\rho(2m/i)^{1/2}] = 1/\alpha_0,
 \end{aligned} \tag{12.6}$$

converts Table 12.1 into Table 12.2.

In this table B , α_0 , α_1 , are defined by (12.6); $m = n + 1$; $-\pi/2$

TABLE 12.2 — LEADING TERMS IN THE ASYMPTOTIC EXPANSIONS FOR $U_n(z)$, $V_n(z)$, $W_n(z)$ WHEN $z = i^{1/2}\rho$ AND $|2n| \gg \rho^2$

Region in m -plane ($m = n + 1$)	$U_n(i^{1/2}\rho)$	$V_n(i^{1/2}\rho)$	$W_n(i^{1/2}\rho)$
Ia	$i^n B\alpha_1$	$-i^{-n} B\alpha_0$	$B(i^{-n}\alpha_0 - i^n\alpha_1)$
II	$B(i^{-n}\alpha_0 + i^n\alpha_1)$	$-i^{-n} B\alpha_0$	$-i^n B\alpha_1$
Ib	$(1 - i^{4n})i^{-n} B\alpha_0$	$i^n B\alpha_1$	$B(i^{3n}\alpha_0 - i^{-n}\alpha_0 - i^n\alpha_1)$
III	$(1 - i^{4n})i^{-n} B\alpha_0$	$B(i^n\alpha_1 - i^{-n}\alpha_0)$	$B(i^{3n}\alpha_0 - i^n\alpha_1)$

$< \arg m \leq 3\pi/2$; and in regions Ia and Ib $\arg m$ is approximately $-\pi/2$ and $3\pi/2$, respectively. Gamma functions may be introduced into the expression for B with the help of (12.4). It may be verified that the functions do not change, except for negligible terms, in crossing over the boundary from Ia to Ib (α_0 and α_1 are interchanged and B is changed by the factor $\exp(-m\pi i)$).

Since the zeros of our functions, regarded as functions of n , occur (asymptotically) when the contributions from two saddle points cancel each other, we may look at Table 12.1 and pick out regions which may possibly contain zeros. Thus, A_0 may equal A_1 along the line $|A_0| = |A_1|$, i.e. very nearly $\operatorname{Re}[f(t_0) - f(t_1)] = 0$, in the m -plane. These lines were discussed in Item 7 of Section 10 and are plotted on the auxiliary w -plane in Fig. 10.3 When plotted on the m -plane the lines appear as shown in Fig. 12.1 The condition $\operatorname{Re}[f(t_0) - f(t_1 \exp(-2\pi i))] = 0$ gives the line $|A_0| \approx |i^{4n} A_1|$ for some of the zeros of $W_n(i^{1/2}\rho)$.

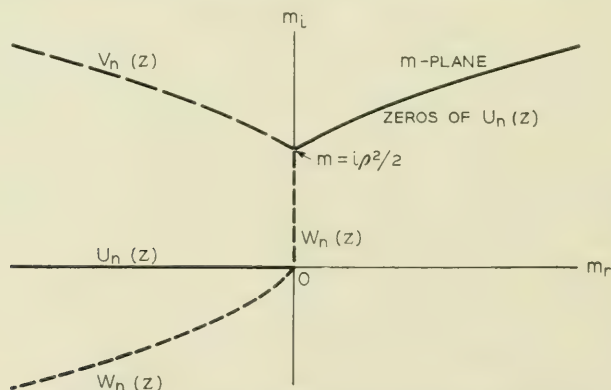


Fig. 12.1 — When $U_n(z)$, $V_n(z)$, and $W_n(z)$ are regarded as functions of n their zeros lie on the lines indicated when $z = i^{1/2}\rho$. The three branches coming out from $m = i\rho^2/2$ are lines along which $|A_0| \approx |A_1|$ and the branch for $W_n(z)$ coming down from $m = 0$ is a line along which $|A_0| \approx |i^{4n} A_1|$ where A_0 and A_1 appear in Table 12.1.

The location of the zeros far out on the lines of Fig. 12.1 may be obtained by writing the appropriate expressions of Table 12.2 as proportional to B times a cosine or sine. Examination of the trigonometrical terms shows that

$$\begin{aligned} U_n(i^{1/2}\rho) &\text{ has zeros at } n \approx 2k + 1 + i^{1/2}4\rho k^{1/2}/\pi, \\ V_n(i^{1/2}\rho) &\text{ has zeros at } n \approx -2k + i^{3/2}4\rho k^{1/2}/\pi, \\ W_n(i^{1/2}\rho) &\text{ has zeros at } n \approx -2k + i^{-1/2}4\rho k^{1/2}/\pi, \end{aligned} \quad (12.7)$$

where k is a large positive integer. Of course, $U_n(z)$ also is zero when n is a negative integer.

So far we have been dealing with $z = i^{1/2}\rho$. Now we consider the case $z = i^{-1/2}\rho$.

Asymptotic expressions which hold when $z = i^{-1/2}\rho$ may be obtained from Table 12.1 by using the relations (9.11) between functions of z and of its complex conjugate z^* . Thus, for example, $V_{a+ib}(i^{-1/2}\rho)$ is equal to the complex conjugate of $W_{a-ib}(i^{1/2}\rho)$. These relations, and relations such as

$$\begin{aligned} [t_0 \text{ for } z = i^{1/2}\rho, n = a - ib]^* &= t_0 \text{ for } z = i^{-1/2}\rho, n = a + ib \\ [f(t_0) \text{ for } z = i^{1/2}\rho, n = a - ib]^* &= f(t_0) \text{ for } z = i^{-1/2}\rho, n = a + ib \end{aligned} \quad (12.8)$$

have been used in constructing Tables 12.3 and 12.4 from Tables 12.1 and 12.2. The interchange of $V_n(z)$ and $W_n(z)$ should be noted. The

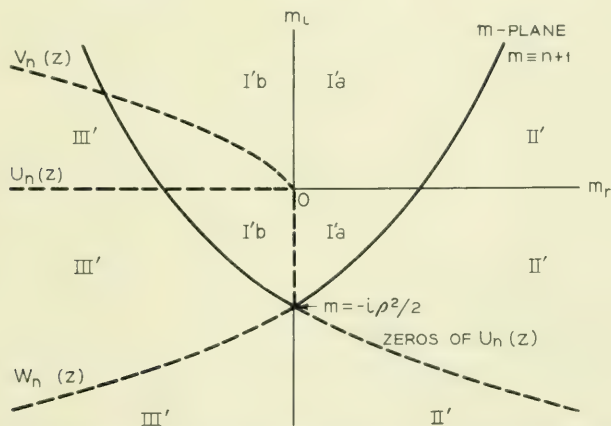


Fig. 12.2 — Regions in the complex m -plane corresponding to different asymptotic expressions when $z = i^{-1/2}\rho$. The lines on which the zeros of the various functions lie are shown by the dashed lines. The corresponding information for $z = i^{1/2}\rho$ is shown on Figs. 11.2 and 12.1.

TABLE 12.3 — LEADING TERMS IN THE ASYMPTOTIC EXPANSIONS FOR $U_n(z)$, $V_n(z)$, $W_n(z)$ WHEN $z = i^{-1/2}\rho$, $\rho > 0$

Region in m -plane $m = n + 1$	$U_n(i^{-1/2}\rho)$	$V_n(i^{-1/2}\rho)$	$W_n(i^{-1/2}\rho)$
Ia'	A'_1	$-A'_0 - A'_1$	A'_0
II'	$A'_1 - A'_0$	$-A'_1$	A'_0
Ib'	$(1 - i^{-4n})A'_1$	$-A'_0 - (1 - i^{-4n})A'_1$	A'_0
III'	$(1 - i^{-4n})A'_1$	$-A'_0 + i^{-4n}A'_1$	$A'_0 - A'_1$

regions in the $m = n + 1$ plane corresponding to the different asymptotic expressions are shown in Fig. 12.2. The boundaries are simply those of Fig. 11.2 reflected in the real m -axis. The lines of zeros are also shown in Fig. 12.2, and are reflections of those of Fig. 12.1 except for the interchange of $V_n(z)$ and $W_n(z)$.

Table 12.3 may also be constructed by returning to the paths of integration shown in Section 11. It may be shown that corresponding to every path of steepest descent for $z = i^{-1/2}\rho$, $n = n_1$ there is another path, obtained from the first by reflection in the real t -axis, which gives the path of steepest descent for $z = i^{-1/2}\rho$, $n = n_1^*$.

The notation used in Table 12.3 is as follows:

$$\begin{aligned}
 z &= i^{-1/2}\rho, m = n + 1, & i &= \exp(i\pi/2), \\
 -3\pi/2 &\leq \arg m < \pi/2, & -3\pi/4 &\leq \arg t_0 < \pi/4, \\
 -3\pi/2 &\leq \arg(-i\rho^2 - 2m) < \pi/2, & -5\pi/4 &\leq \arg t_1 < 3\pi/4, \\
 t_0 &= [i^{-1/2}\rho + (-i\rho^2 - 2m)^{1/2}]/2, & t_1 &= [i^{-1/2}\rho - (-i\rho^2 - 2m)^{1/2}]/2, \\
 A'_0 &= [t_0^{1/2}(-i\rho^2 - 2m)^{-1/4}/(-2i\pi^{1/2})] \exp f(t_0), & (12.9) \\
 A'_1 &= [t_1^{1/2}(-i\rho^2 - 2m)^{-1/4}/2\pi^{1/2}] \exp f(t_1),
 \end{aligned}$$

$$f(t_0) = zt_0 + \frac{m}{2} - m \log t_0 = \frac{m}{2} \left(1 - \log \frac{m}{2} - \log \frac{t_0}{t_1} \right) + i^{-1/2}\rho t_0,$$

$$f(t_1) = zt_1 + \frac{m}{2} - m \log t_1 = \frac{m}{2} \left(1 - \log \frac{m}{2} - \log \frac{t_1}{t_0} \right) + i^{-1/2}\rho t_1.$$

Sometimes it is helpful to use

$$\begin{aligned}
 (2\pi)^{-1/2} \exp \left[\frac{m}{2} \left(1 - \log \frac{m}{2} \right) \right] \\
 \sim \begin{cases} 1/\Gamma \left(\frac{m+1}{2} \right) = 1/\Gamma(1+n/2) & \text{for } -\pi/2 < \arg m < \pi/2, \\ i^m \Gamma \left(\frac{1-m}{2} \right) / 2\pi = i^{n+1} \Gamma(-n/2)/2\pi, & -3\pi/2 < \arg m < -\pi/2. \end{cases} \quad (12.10)
 \end{aligned}$$

TABLE 12.4 — LEADING TERMS IN THE ASYMPTOTIC EXPANSIONS FOR $U_n(z)$, $V_n(z)$, $W_n(z)$ WHEN $z = i^{-1/2}\rho$ AND $|2n| \gg \rho^2$

Region in m -plane $m = n + 1$	$U_n(i^{-1/2}\rho)$	$V_n(i^{-1/2}\rho)$	$W_n(i^{-1/2}\rho)$
Ia'	$i^{-n}B'\alpha'_1$	$B'(i^n\alpha'_0 - i^{-n}\alpha'_1)$	$-i^nB'\alpha'_0$
II'	$B'(i^n\alpha'_0 + i^{-n}\alpha'_1)$	$-i^{-n}B'\alpha'_1$	$-i^nB'\alpha'_0$
Ib'	$(1 - i^{-4n})i^nB'\alpha'_0$	$B'(i^{-3n}\alpha'_0 - i^n\alpha'_1)$	$i^{-n}B'\alpha'_1$
III'	$(1 - i^{-4n})i^nB'\alpha'_0$	$B'(i^{-3n}\alpha'_0 - i^n\alpha'_1)$	$B'(i^{-n}\alpha'_1 - i^n\alpha'_0)$

The notation used in Table 12.4 is as follows:

$$\begin{aligned}
 m &= n + 1, i = \exp(i\pi/2), \quad -3\pi/2 \leq \arg m < \pi/2, \\
 B' &= 2^{-3/2}\pi^{-1/2} \exp\left[\frac{m}{2}\left(1 - \log \frac{m}{2}\right) - \frac{i\rho^2}{2}\right], \\
 \alpha'_0 &= \exp[-\rho(2im)^{1/2}], \\
 \alpha'_1 &= \exp[\rho(2im)^{1/2}] = 1/\alpha'_0,
 \end{aligned}
 \tag{12.11}$$

B' may be expressed in terms of gamma functions with the help of (12.10.).

Approximate expressions for the zeros are given by the complex conjugates of (12.7). For example, if k be a large positive integer such that $2k \gg \rho^2$, the zeros of $W_n(i^{-1/2}\rho)$ are at $n = n(k)$ where $i^{-n}\alpha'_1 = \exp(i\pi k)$ and

$$n(k) \approx -2k + i^{-3/2}4\rho k^{1/2}/\pi - 4i\rho^2/\pi^2.
 \tag{12.12}$$

Here the approximation has been carried out one step further than in (12.7) We also have for the quantities in (7.11)

$$\begin{aligned}
 [\partial W_n(i^{-1/2}\rho)/\partial n]_{n=n(k)} &\approx (-)^{k+1}B'i[\pi - \rho(i/k)^{1/2}], \\
 [i^{2n}V_n(i^{-1/2}\rho)/\sin \pi n]_{n=n(k)} &\approx (-)^{k+1}2B'i.
 \end{aligned}
 \tag{12.13}$$

13. ASYMPTOTIC EXPRESSIONS FOR $U_n(z)$, ETC., WHEN n IS NEAR $z^2/2$

The asymptotic expressions given in Section 12 fail when n is near $z^2/2$. Expressions for the parabolic cylinder functions which hold for this region have been given by Schwid.* More recent studies of this sort, based on differential equations, have been made by T. M. Cherry²⁴ and F. Tricomi.²⁵ Their results suggest the possibility that our expressions

* Reference 20, page 478.
²⁴ Uniform Asymptotic Expansions, J. Lond. Math. Soc., **24**, pp. 121-130, 1949.
 Uniform Asymptotic Formulae for Functions with Transition Points, Am. Math. Soc. Trans., **68**, pp. 224-257, 1950.
²⁵ Equazioni Differenziali, Einaudi, Torino, pp. 301-308, 1948.

for the electromagnetic field which contain Airy integrals may be replaced by more accurate, but also more complicated, expressions. In dealing with our functions we shall work with the integrals and our procedure is somewhat similar to that used by Rydbeck.* First however, we point out that when we write (as suggested by the work of Cherry and Tricomi)

$$y = e^{-z^2/2} T_n(z), \quad ax = z - (2n + 1)^{1/2}, \quad 2(2n + 1)^{1/2} a^3 = 1, \quad (13.1)$$

the differential equation (9.7) for the parabolic cylinder functions goes into

$$\frac{d^2 y}{dx^2} - xy = 2^{-4/3} (2n + 1)^{-2/3} x^2 y. \quad (13.2)$$

The Airy integrals $Ai(x)$ and $Bi(x)$ (and also $Ai[x \exp(\pm i2\pi/3)]$) discussed later in this section are solutions of

$$\frac{d^2 y}{dx^2} - xy = 0, \quad (13.3)$$

and therefore we expect that approximate solutions of (13.2) are given by, for example,

$$y = c_1 Ai(xi^\nu) [1 + O(n^{-2/3})] \quad (13.4)$$

where the $O(n^{-2/3})$ term corresponds to the particular integral of (13.2) when the y on the right hand side is replaced by its approximate value $Ai(xi^\nu)$. Here c_1 is independent of x (or z) but may depend on n , and ν may be 0 or $\pm 4/3$.

Since the labor of computing c_1 is considerable, we shall work out the approximations directly from the integrals.

We shall consider the case $z = i^{1/2} \rho$, $\rho > 0$, first. When $n + 1 = m = m_0 \equiv i\rho^2/2$ the saddle points t_0 and t_1 coincide at $t_2 \equiv i^{1/2} \rho/2$. Consequently only those portions of the paths of steepest descent which lie near t_2 are of importance. This is true even if m is not exactly equal to m_0 . We therefore regard

$$f(t) = -t^2 + 2zt - m \log t \quad (13.5)$$

in (10.1) as a function of the two variables t and m (linear in m) with z fixed at $i^{1/2} \rho$. Expanding (13.5) about $t = t_2$, $m = m_0$ gives

$$f(t) = \frac{z^2}{2} + \left[\frac{m_0}{2} - \frac{m_0}{2} \log \frac{m_0}{2} - \frac{(m - m_0)}{2} \log \frac{m_0}{2} \right] - 4(t - t_2)^3/3z - 2(m - m_0)(t - t_2)/z + \cdots \quad (13.6)$$

* Page 87 of Reference 21 cited on page 478.

where we have used

$$t_2 = z/2 = i^{1/2} \rho/2 = (m_0/2)^{1/2} \quad (13.7)$$

and have arranged the terms within the brackets so that they represent the first two terms in the expansion of

$$\frac{m}{2} - \frac{m}{2} \log \frac{m}{2} \sim \log \left[(2\pi)^{1/2} / \Gamma\left(\frac{m+1}{2}\right) \right] \quad (13.8)$$

about m_0 .

The paths of steepest descent in the t -plane when $m = m_0$ are shown in Fig. 11.6(d). The three branches start out from $t = t_2$ in the directions $\arg(t - t_2) = 15^\circ, 135^\circ$, and -105° . In this section we take the paths of integration to be those of Fig. 11.6(d) even when m is not exactly equal to m_0 . Since we are dealing with asymptotic expressions we may confine our attention to the region around $t = t_2$ where the paths of integration are essentially straight lines [the contributions from $t_2 \exp(-2\pi i)$ are negligible].

When (13.6) is set in the integral

$$(1/2\pi i) \int \exp[f(t)] dt \quad (13.9)$$

we see that the initial directions of the branches are such as to make $(t - t_2)^3/z$ positive ($\arg z = 45^\circ$). Some study of (13.6) and of the Airy integrals we intend to use suggests that we change the variable of integration from t to s and introduce the parameter b where

$$t - t_2 = s(z/4)^{1/3}, \quad b = (m - m_0)(2/z^2)^{1/3} = (m - m_0)/m_0^{1/3}. \quad (13.10)$$

This and (13.6) converts the integral (9.1) for $V_n(i^{1/2}\rho)$ into

$$V_n(i^{1/2}\rho) = \frac{(z/4)^{1/3} e^{z^2/2}}{i(2\pi)^{1/2} \Gamma\left(\frac{m+1}{2}\right)} \int_{-\infty \exp(i2\pi/3)}^{\infty} \exp[-bs - s^3/3 + \dots] ds. \quad (13.11)$$

When we use the Airy integral defined by

$$\begin{aligned} Ai(x) &= \pi^{-1} \int_0^{\infty} \cos(xt + t^3/3) dt \\ &= (i^{1/3}/2\pi) \int_{-\infty \exp(i2\pi/3)}^{\infty} \exp[-i^{2/3}xs - s^3/3] ds, \end{aligned} \quad (13.12)$$

we obtain

$$V_n(i^{1/2}\rho) \sim \frac{(z/4)^{1/3} e^{z^2/2}}{i(2\pi)^{1/2} \Gamma\left(\frac{m+1}{2}\right)} \frac{2\pi}{i^{1/3}} Ai(bi^{2/3}). \quad (13.13)$$

In order to obtain expressions corresponding to (13.13) for $U_n(z)$, $W_n(z)$ we examine Fig. 11.6(d). We have already seen that the limits of integration for s , in the integral (13.11) for $V_n(i^{1/2}\rho)$, are $[\infty \exp(i2\pi/3), \infty]$. In the same way it follows that the limits for $U_n(i^{1/2}\rho)$ and $W_n(i^{1/2}\rho)$ are $[\infty \exp(-i2\pi/3), \infty \exp(i2\pi/3)]$ and $[\infty, \infty \exp(-i2\pi/3)]$, respectively. When we take $s' = s \exp(\mp i2\pi/3)$ as new variables of integration (with the upper sign for $U_n(z)$ and the lower one for $W_n(z)$), the integrals corresponding to (13.11) go into Airy integrals.

We can write our results for $z = i^{1/2}\rho$, when n is close to $i\rho^2/2$, as follows:

$$\begin{aligned} U_n(i^{1/2}\rho) &\sim C i^{1/6} Ai(b i^{6/3}), \\ V_n(i^{1/2}\rho) &\sim C i^{-7/6} Ai(b i^{2/3}), \\ W_n(i^{1/2}\rho) &\sim C i^{9/6} Ai(b i^{-2/3}), \end{aligned} \quad (13.14)$$

where

$$\begin{aligned} C &= (\rho/4)^{1/3} (2\pi)^{1/2} e^{i\rho^2/2} / \Gamma\left(\frac{m+1}{2}\right), \\ b &= (2/\rho^2)^{1/3} (m - i\rho^2/2) i^{-1/3}, \\ i &= \exp(i\pi/2), \quad m = n + 1. \end{aligned} \quad (13.15)$$

The asymptotic expansions whose leading terms are given by (13.14) may be obtained by the method used by F. W. J. Olver²⁶ to study Bessel functions.

$Ai(x)$ and its derivative have been tabulated for positive and negative values of x .^{*} Here we shall use the definitions and results as set forth in Reference 11. These tables and (13.14) enable us to obtain values of $U_n(i^{1/2}\rho)$ along the rays in the m -plane defined by $\arg(m - i\rho^2/2) = \pi/6$ and $-5\pi/6$. Along the $\pi/6$ ray $b i^{6/3}$ is negative. Since the tables show that the zeros of $Ai(x)$ occur when x is negative, it follows that the zeros of $U_n(i^{1/2}\rho)$ occur on the $\pi/6$ ray. In the same way it is seen that the zeros of $V_n(i^{1/2}\rho)$ and $W_n(i^{1/2}\rho)$ occur on the $5\pi/6$ and the $-\pi/2$ rays, respectively. This agrees with Fig. 12.1.

The Airy integral defined by

^{*} Reference 11, page 424.

²⁶ Some New Asymptotic Expansions for Bessel Functions of Large Orders, Proc. Cambridge Phil. Soc., **48**, pp. 414-427, 1952.

$$\begin{aligned}
 Bi(x) &= \frac{i^{-2/3}}{2\pi} \left[\int_{-\infty}^{\infty} \exp(i2\pi/3) \right. \\
 &\quad \left. + \int_{-\infty}^{\infty} \exp(-i2\pi/3) \right] \exp[-i^{-2/3}xs - s^3/3] ds \\
 &= \frac{1}{\pi} \int_0^{\infty} \left[\exp\left(-\frac{t^3}{3} + xt\right) + \sin\left(\frac{t^3}{3} + xt\right) \right] dt
 \end{aligned} \tag{13.16}$$

is also tabulated in Reference 11 where it is shown that

$$\begin{aligned}
 Ai(xi^{4/3}) &= i^{2/3}[Ai(x) - iBi(x)]/2, \\
 Ai(xi^{-4/3}) &= i^{-2/3}[Ai(x) + iBi(x)]/2.
 \end{aligned} \tag{13.17}$$

With the help of these relations we may evaluate the expressions (13.14) for $U_n(i^{1/2}\rho)$, etc., on any one of the six rays

$$\arg(m - i\rho^2/2) = \pm 5\pi/6, \pm \pi/2, \pm \pi/6.$$

When b is a general complex number the expressions (13.14) may be evaluated with the help of the modified Hankel functions $h_1(\alpha)$, $h_2(\alpha)$ tabulated in Reference 27 for complex values of α . The relation needed is

$$\begin{aligned}
 Ai(\alpha) &= \frac{1}{2k} h_1(-\alpha) + \frac{1}{2k^*} h_2(-\alpha), \\
 k &= (12)^{1/6} i^{-1/3}.
 \end{aligned} \tag{13.18}$$

When $|\arg \alpha| < \pi$ we have the asymptotic expansion

$$Ai(\alpha) \sim 2^{-1} \pi^{-1/2} \alpha^{-1/4} (\exp[-(2/3)\alpha^{3/2}](1 - 5/48\alpha^{3/2} + \dots)) \tag{13.19}$$

and when $|\arg(-\alpha)| < 2\pi/3$ we have

$$Ai(\alpha) \sim \pi^{-1/2} (-\alpha)^{-1/4} \sin[(2/3)(-\alpha)^{3/2} + \pi/4]. \tag{13.20}$$

Both of these expansions follow from the discussion of the asymptotic behavior of $h_1(\alpha)$ and $h_2(\alpha)$ given by W. H. Furry and H. A. Arnold.²⁷

Asymptotic expressions for $U_n(i^{-1/2}\rho)$, $\dots$ valid when n is near $-i\rho^2/2$ may be obtained by applying the relations $U_n(z^*) = [U_{n^*}(z)]^*$ $\dots$ given by (9.11) to the expressions (13.14) for $U_n(i^{1/2}\rho)$, $\dots$:

$$\begin{aligned}
 U_n(i^{-1/2}\rho) &\sim C' i^{-1/6} Ai(b' i^{-6/3}), \\
 V_n(i^{-1/2}\rho) &\sim C' i^{-3/6} Ai(b' i^{2/3}), \\
 W_n(i^{-1/2}\rho) &\sim C' i^{-7/6} Ai(b' i^{-2/3}),
 \end{aligned} \tag{13.21}$$

²⁷ Tables of the Modified Hankel Functions of Order One Third and of Their Derivatives, Harvard Univ. Press, 1945.

where

$$\begin{aligned} C' &= (\rho/4)^{1/3} (2\pi)^{1/2} e^{-i\rho^2/2} / \Gamma\left(\frac{m+1}{2}\right), \\ b' &= (2/\rho^2)^{1/3} (m + i\rho^2/2) i^{1/3}, \\ i &= \exp(i\pi/2), \quad m = n + 1. \end{aligned} \quad (13.22)$$

In (13.14) $bi^{6/3} = -b$ and in (13.21) $b'i^{-6/3} = -b'$ since $Ai(\alpha)$ is a single-valued function of α . It is interesting to note that the factor $i^{1/6}$ in the expression for $U_n(i^{1/2}\rho)$ gives the direction of that one of the three paths of steepest descent (in the t -plane) which is *not* traversed in getting $U_n(i^{1/2}\rho)$. The same sort of thing is true for the remaining expressions in (13.14) and (13.21).

The functions

$$'U_n(z) = \exp(z^2/2) \partial[U_n(z) \exp(-z^2/2)] / \partial z,$$

defined by (4.19), may be computed from (13.21) when $z = i^{-1/2}\rho$. We need the relations $\partial/\partial z = i^{1/2}\partial/\partial\rho$ and

$$\begin{aligned} m &= -i\rho^2/2 + b'(\rho^2/2i)^{1/3}, \\ \partial b'/\partial\rho &= (2/3)(2i\rho)^{1/3}(i - m/\rho^2) = i(2i\rho)^{1/3} - 2b'/3\rho, \end{aligned} \quad (13.23)$$

which follow from the definition of b' . When the differentiations are carried out we obtain

$$\begin{aligned} 'U_n(i^{-1/2}\rho) &\sim (2\rho)^{1/3} C' i^{-1/3} Ai'(b'i'^{-6/3}), \\ 'V_n(i^{-1/2}\rho) &\sim (2\rho)^{1/3} C' i^{3/3} Ai'(b'i'^{2/3}), \\ 'W_n(i^{-1/2}\rho) &\sim (2\rho)^{1/3} C' i^{7/3} Ai'(b'i'^{-2/3}), \end{aligned} \quad (13.24)$$

In these expressions the prime on the Airy integral denotes its derivative:

$$Ai'(\alpha) = dAi(\alpha)/d\alpha. \quad (13.25)$$

Abstracts of Bell System Technical Papers*

Not Published in this Journal

AIKENS, A. J.,¹ and C. S. THAELE.²

Control of Noise and Crosstalk on N1 Carrier Systems, A.I.E.E. Trans., Commun. & Electronics, 9, pp. 605-611, Nov., 1953.

BENEDICT, T. S.¹

Microwave Observation of the Collision Frequency of Holes in Germanium, Letter to the Editor, Phys. Rev., 91, pp. 1565-1566, Sept. 15, 1953.

BENNETT, W.¹

Telephone System Applications of Recorded Machine Announcements, Elec. Eng., 72, pp. 975-980, Nov., 1953.

Applications of voice-recording equipment discussed in some detail can be divided into four general groups: Announcements made directly to and providing a service to subscribers, such as weather forecasts and the time of day; announcements to assist subscribers in connection with telephone service, that is, intercept announcements when an individual calls a vacant or disconnected terminal, or emergency announcements if an unusual condition prevents normal service; announcements to expedite service and assist operators in completing calls, including completion of calls from a dial to a non-dial phone, and advising operators of the time delay for completing long distance calls; and specialized announcement or recording services, such as price quotation and ticket reservation.

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ Bell Telephone Laboratories.

² American Telephone and Telegraph Company.

BRIGGS, H. B.,¹ and R. C. FLETCHER.¹

Absorption of Infrared Light by Free Carriers in Germanium, Phys. Rev., **91**, pp. 1342-1346, Sept. 15, 1953.

The absorption of infrared light associated with the presence of free carriers in germanium has been measured by injecting these carriers across a p - n junction at room temperature. The absorption is found to be proportional to the concentration of carriers. The absorption as a function of wavelength shows the same rather sharp maxima previously observed in normal p -type germanium. These bands are found to change with temperature. An explanation of this absorption is offered in terms of a degenerate energy band scheme.

BRIGGS, H. B., see M. TANENBAUM.

CARLITZ, L.,¹ and J. RIORDAN.¹

Congruences for Eulerian Numbers, Duke Math. J., **20**, pp. 339-343, Sept., 1953.

CLARK, M. A., see H. C. MONTGOMERY.

CRABTREE, J.,¹ and B. S. BIGGS.¹

Cracking of Stressed Rubber by Free Radicals, Letter to the Editor, J. Polymer Sci., **11**, pp. 280-281, Sept., 1953.

DICKINSON, F. R., see L. H. MORRIS.

FELCH, E. P.,¹ and J. L. POTTER.¹

Preliminary Development of a Magnetor Current Standard, A.I.E.E. Trans., Commun. & Elec., **9**, pp. 524-531, Nov., 1953.

In the wartime development of the air-borne magnetometer, a method of detecting extremely small changes in magnitudes of magnetic fields was developed. The principle involved was the use of a second-harmonic type of magnetic modulator now known as a magnetor. This instrument can detect changes in magnetic fields in the order of 10^{-5} oersted. A study was made at Rutgers University under the sponsorship of Bell Telephone Laboratories to determine the feasibility of obtaining a standard of current using the magnetor principle.

FLETCHER, R. C., see H. B. BRIGGS.

¹ Bell Telephone Laboratories, Inc.

GOERTZ, M., see H. J. WILLIAMS.

GRAY, M. C.¹

Legendre Functions of Fractional Order, *Quart. Appl. Math.*, **11**, pp. 311-318, Oct., 1953.

GRISDALE, R. O.¹

Formation of Black Carbon, *J. Appl. Phys.*, **24**, pp. 1082-1091, Sept., 1953.

Electron microscopic evidence is presented in support of the hypothesis that black carbon resulting from pyrolysis of gaseous hydrocarbons is produced through the intermediate formation of droplets of complex hydrocarbons. Electron diffraction studies further confirm the hypothesis if, as has been found for particles of carbon blacks, the droplets consist in part of graphitic nuclei arranged with their basal planes tangential to the droplet surface. The carbonization of small solid spherules of highly cross-linked organic polymers is described, and it is shown that the morphology of the carbonization products is wholly analogous to those for pyrolytic carbon and carbon blacks. It is suggested, therefore, that the formation of carbon by the carbonization of solids and by deposition from the gas phase occurs through similar mechanisms and that the two processes are simply two extremes in an infinite series of processes which are all fundamentally alike.

GRISDALE, R. O.¹

Properties of Carbon Contacts, *J. Appl. Phys.*, **24**, pp. 1288-1296, Oct., 1953.

Microphone carbon has been produced by deposition of pyrolytic carbon films over the surfaces of small spherules of silica. The properties of contacts between these spherules are shown to be dependent on the structure and geometry of the carbon surface as determined by electron diffraction and microscopic studies. The graphite-like crystallites in pyrolytic carbon surfaces are more or less preferentially oriented with their basal planes parallel to the surface, and the contact properties depend systematically on the degree of orientation. This is explained in terms of the anisotropy in properties of these crystallites which are closely approximated by those of single crystal graphite which were determined. The contact resistance and its temperature coefficient and the "burning voltage" for carbon contacts are explicable on this basis. However, the microphonic sensitivity of carbon contacts is independent of the surface structure and depends only on the surface geometry.

¹ Bell Telephone Laboratories, Inc.

HARRIS, C. M.¹

Speech Synthesizer, Acoust. Soc. Am., J., **25**, pp. 970-975, Sept., 1953.

"Standardized speech" constructed from building blocks called speech modules has been described; it was synthesized by piecing together bits of magnetic tape containing recorded speech sounds. An electromagnetic device, a "speech module synthesizer," is described here which performs the synthesis automatically. When buttons on a keyboard are pressed, a sequence of corresponding speech modules are automatically recorded on tape exactly in tandem. The modules are selected from a group "stored" on a rotating magnetic drum. The pressing of a button causes an electrical signal corresponding to a module to be reproduced — the electrical switching is so arranged that only one complete module is reproduced for a single button-pressing. This electrical signal is amplified, biased, and then fed into a constantly rotating head which makes contact with stationary magnetic tape and records the signal on it. A 10-ke signal superposed on each stored speech module controls an electromagnetic clutch which (a) measures the length of the recording accurately, and (b) advances the tape at the completion of the recording by the correct amount so that the next recording forms a connected sequence with it. The same module may be used any number of times and in combination with different stored modules, thereby introducing wider experimental control in standardized speech studies. The principle of this type of device could be applied to other classes of problems involving communication of information, as the conversion into speech of typing or of electronically-red printed matter.

HARRIS, C. M.¹

Study of the Building Blocks in Speech, Acoust. Soc. Am., J., **25**, pp. 962-969, Sept., 1953.

Identification of the information-bearing elements of speech is important in applying recent thinking on information theory to speech communication. One way to study this problem is to select groups of building blocks and use them to form standardized speech which then may be evaluated; a method having the advantage of simplicity is described. Individual recordings of the building blocks were made on magnetic tape and then various pieces of tape were joined together to form words. Experiments indicated that speech based upon one building block for each vowel and consonant not only sounds unnatural but is mostly unintelligible because the influences on vowel and consonants are missing which ordinarily occur between adjacent speech sounds. To synthesize speech with reasonable naturalness, the influence factor should be included. Here these influences can be approximated by employing more than one building block to represent each linguistic element and by selecting these blocks properly, taking into account the spectral characteristics of adjacent sounds so as to approximate the

¹ Bell Telephone Laboratories, Inc.

time pattern of the formant structure occurring in ordinary speech. There is no *a priori* method of determining how many building blocks are required to produce intelligible standardized speech. This can only be determined from experiments involving listening tests. Such tests are described.

HOLDAWAY, V. L.¹

Bulb Puncture in Gas Tubes, Electronics, **26**, pp. 208, 210, 212, Nov., 1953.

HOPKINS, I. L.¹

Ferry Reduction and the Activation Energy for Viscous Flow, J. Appl. Phys., **24**, pp. 1300-1304, Oct., 1953.

The relationship proposed by Ferry and his co-workers for the effects of frequency and temperature on the dynamic properties of certain polymers is shown to lead to a method for calculating the activation energy of viscous flow from relaxation, creep, and dynamic test data, the results agreeing with those obtained in steady-state flow. The Ferry reduction explains, and is supported by, observed increases in dynamic modulus and viscosity with increasing temperature.

JONES, T. A.,¹ and W. A. PHELPS.¹

A Level Compensator for Telephotograph Systems, A.I.E.E. Trans. Commun. and Electronics, **9**, pp. 537-541, Nov., 1953.

KARNAUGH, M.¹

Map Method for Synthesis of Combinational Logic Circuits, A.I.E.E. Trans., Commun. and Electronics, **9**, pp. 593-598, disc. pp. 598-599, Nov., 1953.

KOMPFNER, R.,¹ and N. T. WILLIAMS.¹

Backward-Wave Tubes, I.R.E., Proc., **41**, pp. 1602-1611, Nov., 1953.

It has been surmised for some time that a traveling-wave tube in which backward-traveling field components can be excited — such as for instance the "Millman" tube — may oscillate in a backward mode, the RF power emerging at the gun-end of the tube and its frequency depending only on the beam voltage. Experiments with the "Millman" tube show this to be so and oscillations have been observed in the first and second backward spatial-harmonic modes. The latter is excited between 600 and 900 volts, the tube oscillating between 5.9 and 6.4 mm. The former more power-

¹ Bell Telephone Laboratories, Inc.

ful mode is excited between 1,600 and 4,000 volts, the tube tuning continuously between 6.0 and 7.5 mm, thus covering a frequency band of 10,000 mc. Power output of about 10 mw has been measured at 6.4 mm. The tube has also been studied as an amplifier and more than 20-db stable backward gain has been obtained. A simple theory of backward gain and of oscillation starting conditions is given.

LANDER, J. J.¹

Auger Peaks in the Energy Spectra of Secondary Electrons from Various Materials, Phys. Rev., **91**, pp. 1382-1387, Sept. 15, 1953.

The energy spectra of secondary electrons from carbon, beryllium, aluminum, nickel, copper, barium, platinum, and the oxides of beryllium, aluminum, nickel, copper, and barium have been measured with equipment of high stability and sensitivity. Characteristic peaks due to Auger electrons emitted as a result of absorption of a valence electron by an excited x-ray level were observed for all these materials. The peaks exhibit structure which is of some theoretical interest. The structure can be related to the distribution in energy of electrons in the valence band, and it complements that observed in soft x-ray emission work. Since the emission of the Auger electron is not subject to the selection rules governing the emission of x-radiation, additional information can be obtained from the Auger electron energy distribution. Excitation of Auger peaks by a beam of low velocity electrons provides an interesting technique for surface analysis. "Plasma" peaks of the type reported by Ruthemann, and interpreted by Pines and Bohm, were also observed.

LOVELL, G. H., see L. H. MORRIS.

MONTGOMERY, H. C.¹ and M. A. CLARK.¹

Shot Noise in Junction Transistors, Letter to the Editor, J. Appl. Phys., **24**, pp. 1337-1338, Oct., 1953.

MORRIS, L. H.,¹ G. H. LOVELL¹ and F. R. DICKINSON.¹

L3 Coaxial System — Amplifiers, A.I.E.E. Trans., Commun. & Electronics, **9**, pp. 505-517, Nov., 1953 (Monograph 2090).

The line amplifiers for the L3 coaxial system are designed to compensate for the loss of the 4 miles of cable which separate the repeaters; the flat amplifiers are used to compensate for equalizer loss and as transmitting amplifiers. The two types are basically similar, consisting of two feedback amplifiers in tandem, separated by an interamplifier network; in the line amplifier, this network is variable, and is automatically adjusted to com-

¹ Bell Telephone Laboratories, Inc.

pensate for variations in cable temperature, and for small deviations from the nominal 4-mile spacing.

PIERCE, J. R.¹

Spatially Alternating Magnetic Fields for Focusing Low-Voltage Electron Beams, Letter to the Editor, J. Appl. Phys., **24**, p. 1247, Sept., 1953.

PIERCE, J. R.,¹ and L. R. WALKER.¹

"Brillouin Flow" with Thermal Velocities, J. Appl. Phys., **24**, pp. 1328-1330, Oct., 1953.

A type of electron flow in a constant magnetic field is described. The beam of electrons is supposed to be everywhere in thermal equilibrium and the usual Brillouin flow is found when the equilibrium temperature tends to zero. Some considerations are put forward bearing on the choice of a suitable beam temperature in specific problems.

POTTER, J. L., see E. P. FELCH

READ, W. T., JR.¹

Dislocations and Plastic Deformation, Physics Today, **6**, pp. 10-14, Nov., 1953.

Small and exceedingly rare defects in the structure of solids are the "weak links" that determine the strength of materials. The article reviews some fundamental concepts concerning plastic deformation in certain ductile metals.

RIORDAN, J., see L. CARLITZ.

ROMIG, H. G., see R. I. WILKINSON.

SCHNETTLER, F. J., see H. J. WILLIAMS.

SHERWOOD, R. C., see H. J. WILLIAMS.

SHOCKLEY, W.¹

Some Predicted Effects of Temperature Gradients on Diffusion in Crystals, Letter to the Editor, Phys. Rev., **91**, pp. 1563-1564, Sept. 15, 1953.

¹ Bell Telephone Laboratories, Inc.

STEENECK, W. R.¹

N1 Carrier Equipment Design, Commun. Eng., **13**, pp. 26-28, Sept.-Oct., 1953.

Progress in telephone apparatus and in radio equipment design seem to follow converging paths, each contributing something to the other. Bell Laboratories started in the telephone field and adopted radio as an accessory means of transmission. More recently, radio manufacturers have borrowed telephone-circuit techniques for remote controls and multiplexing. The N1 equipment, while it looks more like radio than telephone apparatus, is a most interesting example of economy in manufacture, testing, service, and also in cubic contents. And those gains have been achieved, it should be noted, as part of a program to increase reliability and to reduce the duration of outages.

TANENBAUM, M.¹ and H. B. BRIGGS.¹

Optical Properties of Indium Antimonide, Letter to the Editor, Phys. Rev., **91**, pp. 1561-1562, Sept. 15, 1953.

THAELER, C. S., see A. J. AIKENS.

TIEN, P. K.¹

Traveling-Wave Tube Helix Impedance, I.R.E., Proc., **41**, pp. 1617-1623, Nov., 1953.

The impedance parameter of a circular helix, from which the gain of a helix-type traveling wave amplifier is computed, is investigated for a "Tape-Helix" model. Results obtained in this paper indicate that the impedance has a smaller value than for the "Sheath-Helix" model, and is considerably reduced at larger values of ka , the ratio of the helix circumference to the free space wavelength. A tape helix surrounded by a dielectric medium is analyzed. It is shown that the results obtained from the theory can be used to evaluate the helix impedance for usual types of traveling wave tubes. They have been found to be in agreement with measurements on many tube designs.

WALKER, L. R., see J. R. PIERCE.

WILKINSON, R. I.¹ and H. G. ROMIG.¹

Random Picture Spacing with Multiple Camera Installations, S.M. P.T.E. J., **61**, pp. 605-618, Nov., 1953.

When several high-speed cameras are operated simultaneously, but independently, it is possible that the aggregate of pictures obtained will

¹ Bell Telephone Laboratories, Inc.

satisfactorily cover the space between the pictures provided by any one camera. This paper gives a method for estimating the probability that the longest interval without a picture will not exceed a selected value.

WILLIAMS, H. J.,¹ R. C. SHERWOOD,¹ M. GOERTZ¹ and F. J. SCHNETTLER.¹

Stressed Ferrites Having Rectangular Hysteresis Loops, A.I.E.E. Trans., Commun. & Electronics, **9**, pp. 531-537, Nov., 1953.

A study has been made of the effect of stress on the magnetic properties of ferrites. Rectangular hysteresis loops were obtained by encasing toroidal specimens in plastics which shrink during polymerization. Ferrites having this type of hysteresis loop are useful in magnetic switching and magnetic memory devices.

WILLIAMS, N. T., see R. KOMPFFNER.

WRIGHT, S. B.¹

Higher Frequencies, Aero Digest, 67, pp. 66, 70, 72, Nov., 1953. Spectrum crowding plus new techniques has moved USAF ground-air communications into the ultra-high-frequency bands.

Correction

On page 878 of the July, 1953, issue of the JOURNAL, an error was made in quoting the number of P. H. Richardson's patent in Reference 5. It should have been 2,348,572.

Contributors to this Issue

JOHN T. BANGERT, B.S. in E.E., University of Michigan, 1942; M.S. in E.E., Stevens Institute of Technology, 1947; doctoral studies, Columbia University, 1951-. Bell Telephone Laboratories, 1942-. Following design work on various military systems and test equipment during World War II, Mr. Bangert turned his attention to investigations of fundamental problems in transmission and communication theory. Recently he has been exploring new methods of analysis and synthesis of active and passive networks in the time and frequency domain. Member of Sigma Xi, Tau Beta Pi, Phi Kappa Phi, Eta Kappa Nu, I.R.E., and Association for Computing Machinery.

CLAYTON B. BROWN, B.S. in E.E., Polytechnic Institute of Brooklyn, 1952; Bell Telephone Laboratories, 1937-1940; Western Electric Company, 1940-1943; Bell Telephone Laboratories, 1943-. During World War II he worked on the development of precision potentiometers, later doing test work on the automatic trouble recorder; magnet design and test work on the card translator. He is presently concerned with the planning, analysis, and testing of switching apparatus and solderless wrapped connections.

DONALD F. JOHNSTON, B.S. in E.E., Catholic University of America, 1922; Western Electric Company, 1922-1925; Bell Telephone Laboratories, 1925-. From 1922 to 1924, Mr. Johnston prepared literature describing methods of circuit operation. With the Laboratories he has been concerned with testing and development of toll testing and switching circuits.

J. A. LEWIS, B.S., Worcester Polytechnic Institute, 1944; Sc.M., Brown University, 1948; Ph.D., Brown University, 1950; Corning Glass Works, 1950-1951; Bell Telephone Laboratories, 1951-. With the Laboratories, he has been concerned with mathematical research in theoretical mechanics, piezoelectric crystal vibrations, heat transfer, and stress analysis. Member American Mathematical Society, Sigma Xi.

STEPHEN O. RICE, B.S., Oregon State College, 1929; California Institute of Technology, Graduate Studies, 1929-30 and 1934-35, Bell Telephone Laboratories, 1930 . In his first years at the Laboratories, Mr. Rice was concerned with non-linear circuit theory, with special emphasis on methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigations of telephone transmission theory, including noise theory, and applications of electromagnetic theory. Fellow, I.R.E.

WILLIAM W. RIGROD, B.S. in E.E., Cooper Union Institute of Technology, 1934; M.S. in Engineering, Cornell University, 1941; D.E.E., Polytechnic Institute of Brooklyn, 1950; Westinghouse Electric Corporation, 1941-1951; Bell Telephone Laboratories, 1951 . Since 1935 he has been concerned principally with the design of electron tubes. Member of American Physical Society, Sigma Xi.

MILTON E. TERRY, B.Sc., Acadia, 1937; Ph.D. University of North Carolina, 1951; associate professor in mathematical statistics, Virginia Polytechnic Institute, 1949-1952; Bell Telephone Laboratories, 1952 . With the Laboratories, he is a consulting statistician with the mathematics group, special problems section, working on such projects as sampling, L-3 apparatus development, and the transistor. Member of American Society for Quality Control, Institute of Mathematical Statistics, American Statistical Association, Virginia Academy of Science, Sigma Xi.

C. J. TRUITT, B.S. in Chemistry, Harvard University 1924; New York Telephone Company, Traffic Engineering, 1924-1943. Mr. Truitt transferred to the Long Island area when it was formed in 1927, becoming Trunk Traffic Engineer for that area in 1941. In 1943, he transferred to the toll line engineering group, Operating and Engineering Department, American Telephone and Telegraph Company, where he has since been engaged in developing traffic engineering procedures involving the theory and practice of intertoll trunking.

BOGUMIL M. WOJCIECHOWSKI, Polytechnic Institute of Warsaw, E.E., 1936; Research Staff, Physical Department, Polit. Inst., 1936-1938; Technical Advisor, Polish Stratospheric Board, 1937-1939; National Institute of Telecommunication (Warsaw) 1937-1939; Research Bureau, Industrielle des Téléphones (Paris) 1939-1940; Test Set Development Department, Western Electric Co., 1942 . While working for the Western

Electric Company, Mr. Wojciechowski has been engaged in the development of electrical measuring apparatus. Particularly, his work has been concerned with the industrial precision measuring problems growing out of war and post-war changes in electronic techniques. Some of his contributions are: the development of bridges for temperature coefficient measurements of capacitance; inductance measurements of quartz crystal plates; unbalance measurements of four-branch networks; phase sensitive detectors; and digital frequency combining and selecting systems. Mr. Wojciechowski is a Member of the American Institute of Electrical Engineers and a Senior Member of the Institute of Radio Engineers.

THE BELL SYSTEM

Technical Journal

VOTED TO THE SCIENTIFIC AND ENGINEERING

PECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII

MAY 1954

NUMBER 3

P-N-I-P and N-P-I-N Junction Transistor Triodes J. M. EARLY 517

Arcing of Electrical Contacts in Telephone Switching Circuits.
Part III — Discharge Phenomena on Break of Inductive
Circuits M. M. ATALLA 535

Thickness Measurement and Control in the Manufacture of Poly-
ethylene Cable Sheath W. T. EPPLER 559

Topics in Guided-Wave Propagation Through Gyromagnetic Media
Part I — The Completely Filled Cylindrical Guide
H. SUHL AND L. R. WALKER 579

Coupled Wave Theory and Waveguide Applications S. E. MILLER 661

Theoretical Fundamentals of Pulse Transmission — I E. D. SUNDE 721

Bell System Technical Papers Not Published in this Journal 789

Contributors to this Issue 797

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

S. BRACKEN, *Chairman of the Board,*
Western Electric Company

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. MCNEELY, *Vice President, American Telephone*
and Telegraph Company

EDITORIAL COMMITTEE

E. I. GREEN, *Chairman*

A. J. BUSCH

F. R. LACK

W. H. DOHERTY

W. H. NUNN

G. D. EDWARDS

H. I. ROMNES

J. B. FISK

H. V. SCHMIDT

R. K. HONAMAN

G. N. THAYER

EDITORIAL STAFF

J. D. TERBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXIII

MAY 1954

NUMBER 3

Copyright, 1954, American Telephone and Telegraph Company

P-N-I-P and N-P-I-N Junction Transistor Triodes

By J. M. EARLY

(Manuscript received March 18, 1954)

Theory indicates that the useful frequency range of junction transistor triodes may be extended by a factor of ten by a new structure, the p-n-i-p, which uses a thick collector depletion layer of intrinsic (i-type) semiconductor to reduce greatly the collector capacitance and to increase the collector breakdown voltage. This structure will permit simultaneous achievement of high alpha cutoff frequency, low ohmic base resistance, low collector capacitance, and high collector breakdown voltage. Because of the high breakdown voltages and larger areas per unit capacitance, permissible power dissipations appear much larger than for other high frequency junction types. Theoretical calculations indicate that oscillations at frequencies as high as 3,000 mcps may be possible.

Early exploratory models have verified the basic theory. Progress toward initial design objectives has been encouraging. In general, the observed performance has been consistent with the materials used and the structure achieved. The highest frequency of oscillation obtained to date is 95 mcps. Better performance is expected as technical control of materials and structures is improved.

In the five years since the announcement of the junction transistor by Shockley, great steps have been made in extending its useful frequency range and its power-handling capacity. Recent developments, particularly those which have increased the frequency range,^{2, 12, 13} have brought the performance of practical devices close to ultimate limits

prescribed by structure and material. Further extension of frequency range and, to a lesser degree, of power capability must be sought in new materials or in improved structures. The p-n-i-p* transistor employs a new structure which in theory promises to increase the useful frequency range of junction triodes by a factor of at least ten. In the p-n-i-p, the n region of the base and the p region of the collector are separated by a relatively thick region of i-type (i.e., intrinsic or near-intrinsic, almost free of donor and acceptor centers) semi-conductor. This permits establishment of a thick collector depletion layer at relatively low voltages, thus producing low collector capacitance and several other desirable features.

The advantages of the new structure may be seen by study of the limitations of previous triode structures. In general, high frequency performance of conventional units, such as p-n-p alloy¹ transistors, is improved by making the base region thinner to increase the alpha cutoff frequency (f_α), by using lower resistivity base material to reduce the ohmic base resistance (r_b'), and by decreasing the area of emitter and collector junctions to reduce the collector capacitance (C_c). These equivalent circuit parameters are of nearly equal importance as may be seen from the gain-bandwidth expression discussed below.

The design changes required to improve the parameters involve conflicts, and compromises are necessary. For example, the decrease of base thickness which increases alpha cutoff frequency also increases (less rapidly) the ohmic base resistance.† The decrease in base resistivity which reduces base resistance also increases (again, less rapidly) the collector capacitance and decreases the collector breakdown voltage, thus decreasing power capacity. The reduction of junction area which decreases collector capacitance reduces the current rating and thereby the possible power rating. For transistors having circular electrodes, it may also increase the ohmic base resistance.

For these reasons, conventional junction triodes designed for high frequency application tend to be very small and to have very low voltage, current, and power ratings. Ultimately, the decrease of collector reverse breakdown voltage sets a lower limit to usable base resistivity and thereby to the thickness of the collector depletion region. This sets a lower limit on base region thickness, since average base layer thickness should be two or more times depletion layer thickness. For base layers thinner than this, irregularities in thickness or in impurity distribution may permit the depletion layer to contact the emitter, producing the

* And its homologue, the n-p-i-n.

† In the junction triode, this increase of base resistance is overcome by crowding the minority carrier emission close to one of the base contacts, thus producing low ohmic base resistance. See Reference 2.

ac collector-to-emitter short circuit effect called "electrical punch-through." Lower limits of junction areas are set by desired operating currents and by mechanical reasons. Diminishing returns are reached for structures a few mils in diameter and a fraction of a mil thick.

To facilitate comparisons, the limitations described qualitatively above have been interpreted quantitatively in terms of a gain-band figure of merit,*

$$\Gamma_0 \cdot B^2 = \frac{f_\alpha}{25r_b' C_e}; \quad (1.1)$$

in which Γ_0 is low frequency available power gain in the common emitter connection† and B is the frequency at which the gain is 3 db down from its low frequency value. A reasonable upper limit on this (power gain) $\times$ (bandwidth-squared) product is 4×10^{16} , which indicates that a 0–10 mcps video gain of 26 db may be obtained by improvement of conventional triode structures.

The same figure of merit, for a p-n-i-p of equal junction area, is approximately 10^{19} . Calculation shows that units may be designed to produce 10 db or more gain at 1,000 mcps. Although many of its operating principles are similar to those of the p-n-p and the n-p-n, the p-n-i-p differs from the earlier triodes in that low collector capacitance is obtained by means of a thick collector depletion (space-charge) layer of intrinsic semi-conductor. The section view of a p-n-i-p in Fig. 1 illustrates its major features. The wide depletion layer (electric field region) produces small collector capacitance (C_e) and gives a high reverse breakdown voltage, while the very thin base region of low resistivity gives simultaneously a low ohmic base resistance (r_b') and a very high alpha cutoff frequency (f_α). The design with four regions, emitter, base, depletion layer, and collector, increases the (power gain) $\times$ (band-squared) figure of merit ($f_\alpha/25r_b' C_e$) about two decades, thus increasing the useful frequency range about one decade.

The thick collector depletion layer of intrinsic or near-intrinsic semi-conductor provides advantages in addition to the reduction of the collector capacitance. Because base layer resistivity does not limit collector breakdown voltage as it does in previous structures, much lower base resistivities may be used, thus producing lower ohmic base resistances. Furthermore, the thick depletion region makes the structure much more rugged for very high alpha cutoff units since the very thin base layer is

* This figure of merit is essentially identical with one described by R. L. Pri-chard at the A.I.E.E. Winter Meeting in New York City, Jan. 22, 1954.

† It is assumed that the input terminals of the transistor are shunted by an external resistance which determines the input impedance and therefore the band-width. Power gain decreases approximately 6 db per octave at frequencies greater than B .

a surface layer on an 0.5–2.0 mil intrinsic layer, rather than a thin and unsupported web.

When operating biases are applied to a p-n-i-p transistor, holes injected at the forward biased emitter diode diffuse across the n region of the base then drift at high velocities through the field region to the reverse biased collector p region just as in a PNP transistor. However, in the p-n-i-p, the drift transit time through the collector field is comparable to the diffusion transit time through the base and contributes to phase shift of the short-circuit current-transfer ratio, α . In addition, the emitter depletion layer capacitance, C_{Te} , which is unimportant in previous triodes, is relatively large in the p-n-i-p and degrades performance at very high and microwave frequencies by providing a low impedance shunt around the emitter junction.

The details of structure and operation, design theory, a comparison of p-n-p and p-n-i-p units and some experimental results are discussed in the following sections. The concluding summary reviews the theoretical and experimental work.

STRUCTURE AND OPERATION

Impurity Distribution

In general, device characteristics depend on structure and on operating conditions. However, structure is more basic than operating conditions. The spatial distribution of fixed charge centers (donors and ac-

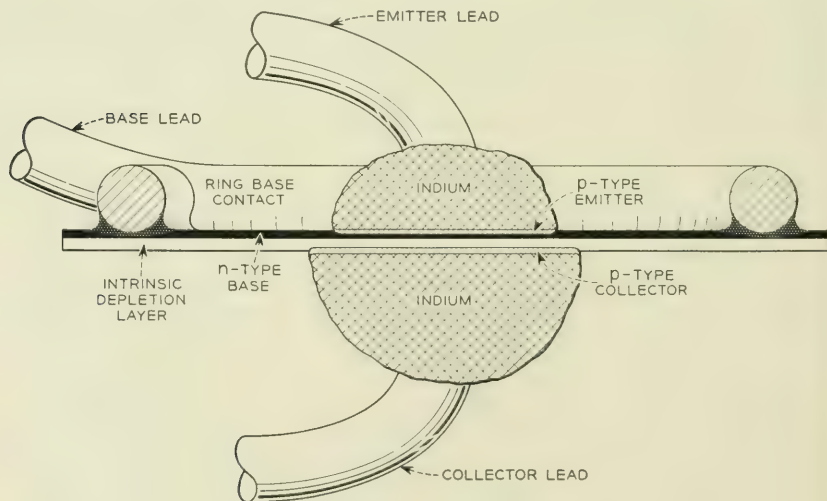


Fig. 1 — Sectional view of a p-n-i-p transistor.

ceptors) is the fundamental structural characteristic of the junction transistor. Fig. 2(a) shows an impurity density profile for a p-n-i-p along an axial line running through emitter, base, collector space-charge layer, and collector. Similar profiles for step junction (alloy) and graded junction (grown crystal) p-n-p's are shown in Figs. 2(b) and (c).

The emitter and collector regions of the p-n-i-p have very high impurity concentrations (low resistivities), while the impurity density in the base is moderately high and the depletion layer is almost free of impurities. The high acceptor density in the emitter forces most of the emitter current to flow as holes, giving an injection ratio (γ) close to unity. The high density in the collector gives a low collector body resistance and fixes the position of one face of the collector depletion layer.

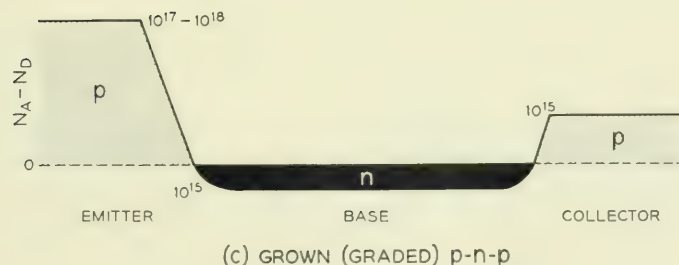
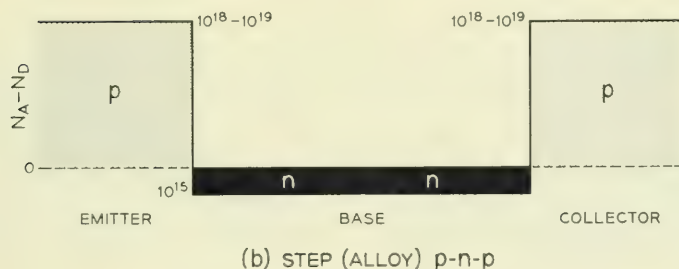
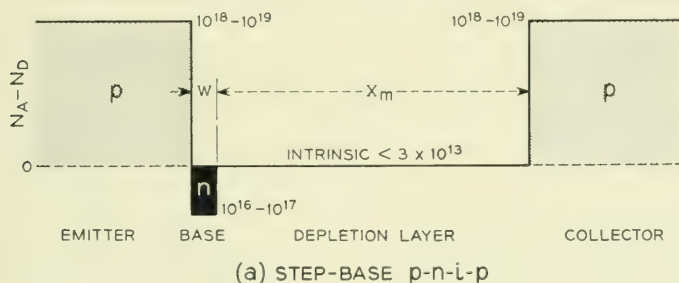


Fig. 2 — Impurity density profiles.

The high donor concentration in the base region leads to low ohmic base resistance (r_b') and fixes the position of the base face of the depletion layer. In the depletion layer, the concentration of impurities is so low that the field region (space-charge layer) extends from the n-type base to the p-type collector at low voltages.

Depletion Layer

The properties of the depletion layer which are important at high frequencies are the capacitance across it (C_c) and the carrier transit time through it (τ_c). These are determined primarily by the impurity density, the thickness of the region, and the base-to-collector voltage. Potential and field distributions in the depletion layer for both small and

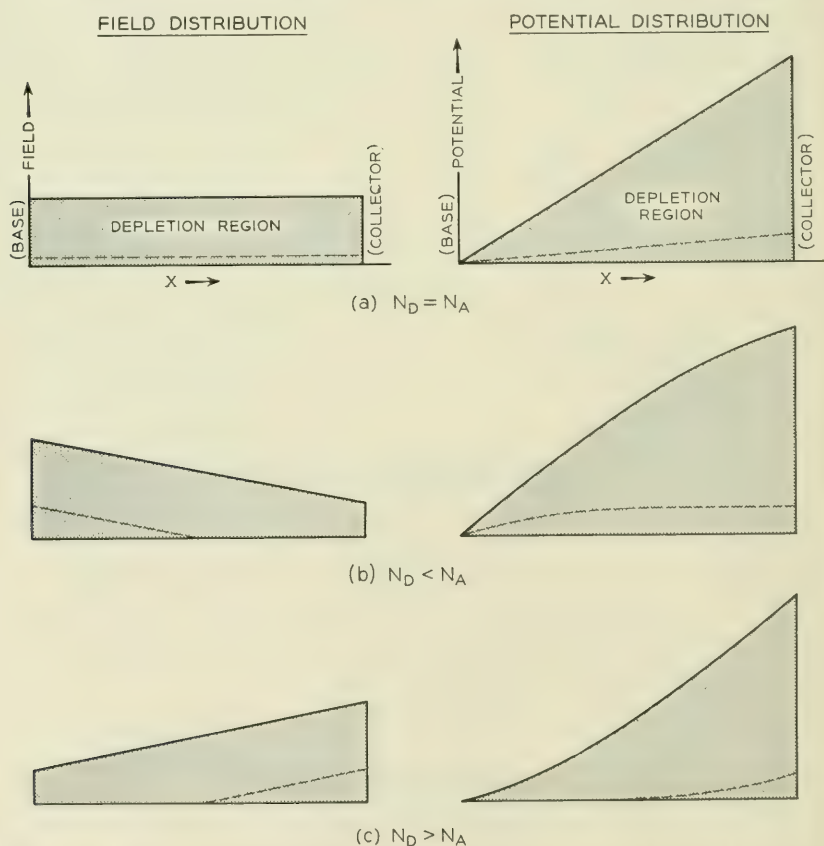


Fig. 3 — Field and potential distributions in depletion region of p-n-i-p transistor.

typical applied voltages are shown in Fig. 3 for p-n-i-p structures in which the depletion layer contains no net impurities (a), a small acceptor dominance (b), and a small donor dominance (c). When collector voltage is increased from zero, the space charge layer thickens until it extends from base to collector. Further increase of voltage simply increases the field strength in the region, without significant further increase in its thickness.

The capacitance initially changes inversely as the square root of collector potential, but becomes constant when depletion region thickness becomes constant. The time required for holes to drift from base to collector decreases with increase of depletion region field until scattering-limited carrier velocities are reached (about 5×10^6 cm/sec for holes, at 10,000 volts/cm).¹⁰ It should be noted that normal operation does not occur until the depletion layer extends from base to collector (particularly if the depletion region is slightly n-type so that effective base thickness is large at low collector voltages, see Fig. 3(c)). The breakdown voltage of the collector is very high,* since the field strength in the depletion region is relatively uniform by comparison with that in older types of units, the region is wide, and strong fields are required to produce carrier multiplication.

Base Region

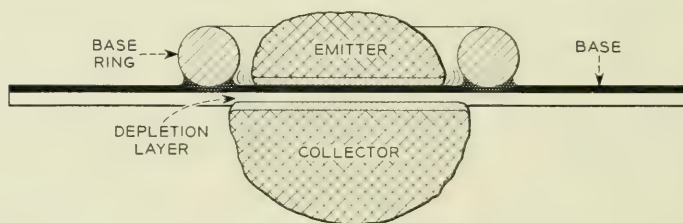
Base region design seeks the conflicting objectives of short diffusion transit time, requiring a thin region, and low ohmic base resistance, requiring a thick region. In practice, the region is made as thin as feasible, but of low resistivity material, and base contact geometry is chosen to minimize the ohmic resistance. In the p-n-i-p, very low base resistivity is practical, because the collector breakdown potential is fixed by the thickness of the intrinsic depletion layer rather than by the base resistivity as in fused junction p-n-p's.

The large donor density in the base region together with the very high frequencies of operation make the emitter depletion layer capacitance (C_{Te}) both larger and more important than in previous transistors. In order to reduce this capacitance, the emitter junction area is made small, thus leading to emitter current densities of 1 to 100 amperes/cm². In general, as the dc current density is increased, the minimum dc collector voltage must also be increased in order to preserve

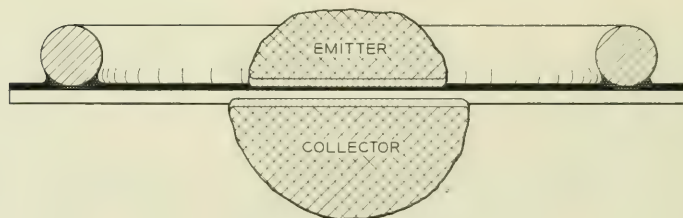
* An avalanche mechanism similar to a Townsend discharge in gases is now believed responsible for reverse voltage breakdown in junction structures. See Reference 3.

emission-limited current flow. Insufficient voltage may result in space-charge limited operation.^{4, 5}

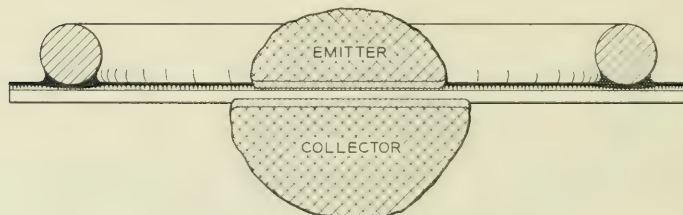
Three structures which may be used to obtain low ohmic base resistance are shown in Fig. 4. Obviously, the base contact ring may be placed arbitrarily close to the emitter, as in Fig. 4(a), so that the base resistance is that of the region beneath the emitter. Since this is somewhat difficult, the ring may be placed at a distance from the emitter, and the emitter imbedded in the base n-region as in Fig. 4(b), reducing the resistance between the emitter periphery and the base ring at only a small cost in alpha cutoff frequency. In addition, as shown in Fig. 4(c), the n-region used may be of graded resistivity such as results from impurity diffusion from the surface. The large impurity concentration at the surface minimizes both edge emission and radial base resistance.



(a) CLOSE-SPACED RING



(b) IMBEDDED EMITTER



(c) IMBEDDED EMITTER - DIFFUSED SURFACE LAYER

Fig. 4 — Low-base resistance structures.

These advantages are, however, balanced in part by an increase in the emitter depletion region capacitance associated with the low resistivity base material.

DESIGN THEORY

General

The principal objectives in the initial p-n-i-p design have been high alpha cutoff frequency, low collector capacitance, and low ohmic base resistance. The equivalent circuit employed is shown in Fig. 5. The output and feedback admittances which are important in earlier junc-

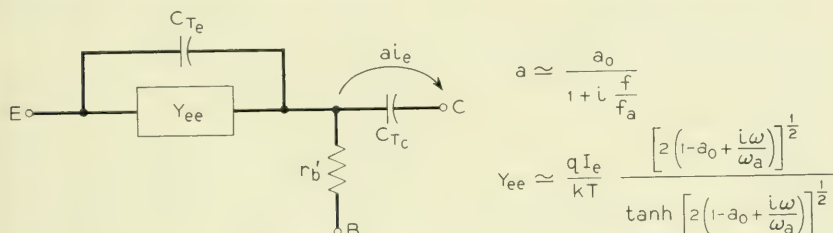


Fig. 5 — Equivalent circuit of the p-n-i-p transistor.

tion triodes are omitted, since the space charge layer widening factor (H_{12} or $\mu_{ec}, \frac{kT}{qw} \frac{\partial w}{\partial V_e}$) is very small.^{6, 7} The transfer admittance is shown as a current generator (ai_e) with cutoff frequency ($|a^2| \sim 3$ db down) of f_a because this gives explicit recognition to base region diffusion transit time τ_b and allows it to be combined with space charge layer transit time τ_c .

Emitter Region Design

Emitter region acceptor concentration should be very large ($10^{18} - 10^{19}$ atoms/cc) in order to keep the injection ratio γ close to unity at both low and high frequencies.⁸ At low frequencies, γ is determined by emitter resistivity and carrier life path or diffusion length, base resistivity and width, as

$$\gamma = \gamma_0 = \frac{1}{1 + \frac{\sigma_b W}{\sigma_e L_{ne}}}$$

At high frequencies, γ is determined by the ratio of acceptor density in the emitter to donor density in the base as

$$\gamma_{\text{hf}} = \frac{1}{1 + \frac{N_{\text{Db}}}{N_{\text{Ae}}} \sqrt{\frac{D_n}{D_p}}}$$

Obviously, since the effective donor density in the base must be large to give low ohmic base resistance, the effective acceptor density in the emitter must be even larger if high frequency γ is to be close to unity.

Base Region Design

Base region thickness, w , and the diffusion constant, D_p , determine the diffusion transit time for holes from injection by the emitter to collection by the field of the depletion layer.

$$\tau_b = \frac{w^2}{2D_p}. \quad (1)$$

For circular electrodes, which are useful, easily made, and easily analyzed, the ohmic base resistance for the active region of the base between emitter and collector depends on base resistivity, ρ_b , and base thickness as follows:

$$r_b' = \frac{\rho_b}{8\pi w} = \frac{1}{q\mu_n N_D 8\pi w}. \quad (2)$$

If w is made small, r_b' can be reduced only by making N_D large. Although large reductions in r_b' can be made, increasing N_D is ultimately a self-defeating procedure for several reasons: as N_D is increased both D_p and the electron mobility, μ_n , decrease, thus increasing hole transit time and also partially off-setting the reduction in r_b' by N_D . In addition, the capacitance of the emitter depletion region varies approximately as $N_D^{1/2}$, thus diverting more ac emitter current from hole injection. This capacitance is

$$C_{\text{Te}} = \kappa\epsilon_0 \left(\frac{qN_d}{2\kappa\epsilon_0 V_e'} \right)^{1/2} A_e, \quad (3)$$

where V_e' is the average electrostatic potential across the emitter depletion layer. Equations (1) to (3) show the conflicts which necessarily arise in base region design for very high frequencies. The limiting design combines very small w , large N_D , small emitter area A_e , and relatively

large dc emitter current I_e so that the minority carrier emitter admittance y_{ee} is at least of the order of magnitude of $j\omega C_{Te}$. Total emitter admittance is

$$y_{ee} + j\omega C_{Te} = \frac{\frac{qI_e}{kT} (1 + j\omega\tau)^{1/2} \coth \left[\frac{(1 + j\omega\tau)w^2}{(D\tau)} \right]^{1/2}}{\coth \left(\frac{w^2}{D\tau} \right)^{1/2}} + j\omega C_{Te} \quad (4)$$

Depletion Layer Design

As mentioned previously, the most important characteristics of the depletion layer in the p-n-i-p are the transit time for holes, τ_c , and the capacitance, C_{Te} . The minimum voltage for normal operation, V_{min} , and maximum or breakdown voltage, V_{max} , are also significant.

The minimum voltage for "normal" operation is reached when the electric field between the n-type base and p-type collector is strong enough so that the holes drift at their limiting velocity of 5×10^6 cm/sec.* The collector to base voltage required for normal operation is the product of the minimum field strength for the limiting velocity and the thickness of the depletion layer and is given by

$$V_{min} = 10,000 x_m \quad (5)$$

in which x_m is depletion layer thickness in cm. The maximum field obtainable before reverse voltage breakdown is not known exactly, but is in practice near 100,000 volts/cm, so that

$$V_{max} \simeq 100,000 x_m. \quad (6)$$

Depletion layer capacitance is nearly independent of collector voltage in normal operation and is inversely proportional to layer thickness.

$$C_{Te} = \frac{\kappa\epsilon_0 A_c}{x_m} \quad (7)$$

Transit time for holes increases directly with layer thickness, however, being

$$\tau_c = \frac{x_m}{5 \times 10^6} \quad (8)$$

Since increase of τ_c decreases the alpha cutoff frequency f_α , the choice

* At lower field strengths, the transit time for holes is longer, giving a lower alpha cutoff frequency. The "normal" is the best, rather than the only possible, operating condition.

of x_m is a design balance between C_c and f_α , with any desire for low voltage operation weighting the scales toward smaller x_m .

As collector voltage and therefore field strength is reduced below that required for normal operation, transit time is increased because of the reduced drift velocity. In addition, the holes in transit interact with the ac field of the layer, thus increasing the output conductance g_{cc} . Further, a larger density of holes in the layer is required to carry the same current, disturbing the field distribution. If the voltage is reduced greatly, space-charge limited emission may occur,^{4, 5} producing much longer effective transit times.

If output voltage is reduced sufficiently, the collector field will not extend all the way from base to collector. If the layer is somewhat n-type, the field region collapses toward the p-region of collector. If it is somewhat p-type, the field collapses toward the n-region of the base. The latter arrangement has the advantage that f_α is less drastically reduced. Further, in normal operation, the negatively charged acceptor atoms of a slightly p-type layer will neutralize the charge of the holes in transit, thus making the field more nearly constant from collector to base. The effects of low voltage on the collector field distribution are indicated approximately by the dashed lines of Fig. 4.*

Collector Region Design

Acceptor concentration in the collector should be large for several reasons. This gives a low collector body resistance, which virtually eliminates internal series loading of the collector, and it aids operation by fixing the position of the collector edge of the depletion layer. The advantages obtained may be seen by considering a unit in which the collector body is made somewhat p-type and a collector contact is attached at some distance from the depletion layer. If 10 ohm-cm p-material is used for the collector body and a collector contact fastened 2.5 mils from the collector resistance of 250-500 ohms will result. In addition, because of the weak drift field at the collector edge of the depletion layer, the hole transit time is about twice that for a true p-n-i-p.

Alpha Cutoff Frequency

A current transmission cutoff frequency f_α for the p-n-i-p is given approximately by†

* The field distributions occurring in an intrinsic depletion layer at low field strengths have been discussed in Reference 11.

† It is assumed that alpha is given by $\alpha = \alpha_0(1 + jf/f_\alpha)$. Equation (9) represents the phase of this expression quite well, but the amplitude rather poorly.

$$f_{\alpha} = \frac{1}{2\pi(\tau_b + \tau_c/2)} \quad (9)$$

Equation (9) implies (correctly) that the delay time for total current passing through the depletion layer is about one-half the transit time for the carriers. This results from the induction of charge on the base and collector electrodes by the carriers in transit. If $\varphi = \omega\tau_c$ is carrier transit angle and $J_c = e^{j\omega t}$ is the *conduction* current of holes entering the depletion layer from the base, the *total* current entering the depletion layer from the base can be shown to be

$$J = e^{j\omega t} \left(\frac{1 - e^{-j\phi}}{j\phi} \right) \quad (10)$$

which reduces for small φ to

$$J \simeq e^{j\omega t - j\phi/2}$$

It may be noted that the total current J of equation (3.6-2), when written in the form $J_{\max} \angle \theta$ in which θ is the phase shift of the total current with respect to the conduction current entering from the base, is approximately $0.973 \angle -22.5^\circ$ for $\varphi = 45^\circ$, $0.901 \angle -45^\circ$ for $\varphi = 90^\circ$, and $0.636 \angle -90^\circ$ for $\varphi = 180^\circ$.

DESIGN COMPARISON

General

Comparison of figures of merit is the best, albeit unsatisfactory, means for comparative evaluation of devices. For junction transistors, one non-controversial figure of merit is established — the noise figure. Two transmission figures of merit for junction transistors are suggested at the bottom of Table I. It should be pointed out that the p-n-i-p figures are for theoretical design possibilities, some features of which have already been realized experimentally.

The Units

Table I gives parameters of interest for several types of transistors. Structural, material, and electrical parameters for the Bell Telephone Laboratories' developmental M1778 p-n-p unit are averages for large numbers of units. The electrical parameters of the plated-contact transistor recently announced by Philco were taken from a talk by W. H. Forster before the Philadelphia I.R.E., Dec. 3, 1953.¹² The structural and material parameters have been estimated. The p-n-i-p structures

TABLE I — TRANSISTOR DESIGNS

	M1778	Philco	P-N-I-P (Calculated Values)		
			No. 1	No. 2	No. 3
w_b — mils	1.0	0.2	0.13	0.8	0.04
ρ_b — ohm cm	1.5	0.5	0.14	0.05	0.02
dia_e — mils	15	4	10	6	5
dia_c — mils	30	6	15	8	5
x_m — mils	0.1	0.05	0.63	0.36	0.7
N_b — atoms/cc	10^{15}	3.5×10^{15}	1.4×10^{16}	4.2×10^{16}	1.2×10^{17}
f_α — mcps	2.0	55	100	200	360 (600)*
r_b' — ohms	50	65	34	20	16
C_c — mmf	25	2.5	1.0	0.5	0.1
$\omega_a C_c$ — mhos	—	—	0.023	0.038	0.102
C_e — mmf	—	—	36	22	27
$\omega_a r_b' C_c$	0.0157	0.056	0.0214	0.0126	0.0060
$(f_\alpha/25 r_b' C_c)^{1/2}$ — mcps	8	115	340	900	3000

* First value calculated by Equation (9); second value is for diffusion through base n-region only (i.e., $\tau_c = 0$).

and materials were assumed and electrical parameters were calculated from them by the Equations (1) to (11). Mobilities measured for low resistivities by M. B. Prince⁹ were used in the calculations.

Figures of Merit

The last row of Table I gives $(f_\alpha/25r_b'C_c)^{1/2}$, which was discussed previously as a gain-bandwidth figure of merit for a broad band common emitter amplifier. It is also related to the maximum frequency at which reliable oscillations may be obtained. The figure of merit $\omega_a r_b' C_c$ is the open circuit voltage feedback ratio at the alpha cutoff frequency and gives some indication of the balance between the two time constants, $1/\omega_a$ and $r_b' C_c$. It is also approximately the ratio of input impedance to output impedance in a common emitter broadband amplifier at high frequencies.

Comments

It should be noted that the emitter depletion layer capacitance is significant in all the p-n-i-p designs and that barrier transit time reduces alpha cutoff frequency some forty per cent in the highest frequency design. Despite this, it is probable that p-n-i-p or n-p-i-n germanium junction triodes will serve as oscillators and perhaps amplifiers at frequencies as high as 3,000 mcps.

EXPLORATORY MODELS

Objectives

While the p-n-i-p transistor will be useful for high voltage and high power operation, our exploratory development work has been directed toward good performance at very high frequencies. The initial electrical objectives set were those of p-n-i-p No. 1 of Table I: $f_\alpha = 100$ mcps, $C_e < 1.0$ mmf, and $r_b' = 34$ ohms. The base thickness of 0.13 mil and base resistivity of 0.14 ohm-cm are the critical structural parameters.

Fabrication

Although p-n-i-p's might conceivably be built in a single operation, one procedure used has two major parts. The first is the production and evaluation of 2-mil thick wafers of intrinsic germanium with a skin or surface layer of 0.1–1.0 ohm-cm n-germanium 0.3–0.5 mils thick. The second step is the alloying of collector, emitter, and base electrodes to these wafers.

Wafers with n-type skins have been made by three methods. Intrinsic crystals growing from a melt by the Teal-Little technique have been doped with arsenic, grown for a few seconds longer (another 0.5–1.0 mils), and snatched mechanically from the melt. The resulting crystal surface has a mirror finish and is relatively flat. N-type skin layers have also been produced by alloying the wafer surfaces with lead-arsenic and lead-antimony mixtures and by the diffusion of arsenic into wafer surfaces.

Collector and emitter electrodes are alloyed by the indium germanium process with times, temperatures, and quantities of indium selected to give desired alloying depths. Ring-base connections of antimony and gold plated kovar have been used.

Measurements

Progress toward the initial design objectives mentioned previously has been encouraging. The predicted behavior has been verified semi-quantitatively. The capacitance of a 15-mil diameter collector is usually less than 1.0 mmf at $V_e = -25$ volts as predicted in design No. 1 of Table 1. Ohmic base resistances generally less than 50 and as low as 5 ohms have been measured. However, the highest alpha cutoff frequency obtained as yet is 25 mcps. This has been limited primarily by the thickness of the base layer. At present this is of the order of 0.30 mils so that an alpha cutoff frequency of 25 mcps is about what would be predicted. Further development of the technology of fabrication seems reasonably

straight-forward at least to the design objectives of No. 1 and No. 2 of Table I.

The best unit measured to date showed $\alpha_0 > 0.96$, $f_\alpha \simeq 25$ mcps, $r_b' \simeq 60$ ohms, and $C_c \simeq 1.8$ mmf. These values agree quite well with those expected from the resistivities and layer thicknesses employed. The unit oscillated at 95 mcps with $V_c = -30$, $I_e = 1.0$ ma. Connected in a common emitter video amplifier working from a 75-ohm generator impedance into a load resistance of 2,150 ohms shunted by 5 mmf of capacitance, this unit produced a power gain of 23 db at 500 kc, falling to 20 db at 3 mcps and 15 db at 10 mcps.* In an uncompensated common emitter tuned circuit, this unit gave 20.5 db at 10 mcps with 3 mcps bandwidth between the three db points.* It has been operated with a collector voltage of -90 volts.

SUMMARY

The designed elimination of donors and acceptors from a thick collector depletion layer introduces a new design variable in junction transistor triodes. The new structure (p-n-i-p or n-p-i-n) is believed capable of development into the microwave frequency range. Several factors which were of second order importance in p-n-p and n-p-n units such as emitter depletion layer capacitance and collector transit times become significant in limiting ultimate performance. The thick depletion layer permits operation at higher voltages than were previously possible in any but low frequency units.

Moderately good results have been obtained already. Units having 10 mil emitter diameter, 15 mil collector diameter have produced stable gains without compensation of 20.5 db at 10 mcps and have oscillated at 95 mcps.

The junction transistor now promises to be a serious competitor to high vacuum triodes over a much larger range of frequencies and power levels than before.

ACKNOWLEDGMENTS

J. A. Morton and R. M. Ryder have strongly supported and encouraged this work. J. W. Peterson and W. C. Hittinger have collaborated in and contributed to the experimental studies. The models constructed and tested are the products of the persistent efforts and many useful suggestions of J. A. Wenger, J. McGlasson, and L. P. Meola. Many others, particularly those engaged in semiconductor materials research

* These measurements were made by L. G. Schimpf.

and development, have also assisted us. Discussions with colleagues have been most helpful in preparation of this report.

REFERENCES

1. J. S. Saby, Fused Impurity p-n-p Junction Transistors, I.R.E. Proc., **40**, pp. 1358-1360, Nov., 1952.
2. R. L. Wallace, Jr., L. G. Schimpf and E. Dickten, A Junction Transistor Tetrode for High-Frequency Use, I.R.E. Proc., **40**, pp. 1395-1400, Nov., 1952.
3. K. G. McKay, and K. B. McAfee, Electron Multiplication in Silicon and Germanium, Phys. Rev., **91**, pp. 1079-1084, Sept. 1, 1953.
4. W. Shockley, and R. C. Prim, Space-Charge Limited Emission in Semiconductors, Phys. Rev., **90**, pp. 753-758, June 1, 1953.
5. G. C. Dacey, Space-Charge Limited Hole Current in Germanium, Phys. Rev. **90**, pp. 759-763, June 1, 1953.
6. J. M. Early, Effects of Space-Charge Layer Widening in Junction Transistors, I.R.E. Proc. **40**, pp. 1401-1406, Nov., 1953.
7. J. M. Early, Design Theory of Junction Transistors, B. S. T. J., **32**, pp. 1271-1312, Nov., 1953.
8. W. Shockley, M. Sparks and G. K. Teal, The p-n Junction Transistors, Phys. Rev., **83**, p. 151, July, 1951. See also Reference 7. Nov. 1953, op cit.
9. M. B. Prince, Drift Mobilities in Semiconductors. I. Germanium. Phys. Rev., **92**, pp. 681-687, Nov. 1, 1953.
10. E. J. Ryder, Mobilities of Holes and Electrons in High Electric Fields, Phys. Rev., **90**, p. 766, June, 1953.
11. R. C. Prim, D. C. Field in a Swept Intrinsic Semiconductor, B. S. T. J., **32**, pp. 665-694, May, 1953.
12. W. E. Bradley, et al, The Surface Barrier Transistor, I.R.E. Proc., **41**, pp. 1702-1720, Dec., 1953.
13. C. W. Mueller and J. I. Pankove, A p-n-p Alloy Triode Transistor for Radio Frequency Amplification, RCA Review, **14**, pp. 586-598, December, 1953.

Arcing of Electrical Contacts in Telephone Switching Circuits

Part III—Discharge Phenomena on Break of Inductive Circuits

By M. M. ATALLA

(Manuscript received November 16, 1953)

This is a presentation of a study of the discharge phenomena occurring between contacts on break of an inductive load. The main objectives are: (1) to forward some detailed explanations of the main components of a break transient in terms of basic conduction and emission processes, and (2) to establish the conditions that determine the nature of the transients. The study covered the following: (1) occurrence of interrupted and steady arcs, (2) initiation of reversed arcs in one breakdown, (3) arc initiation under dynamic conditions, (4) initiation and maintenance of glow discharge, and (5) glow-arc transitions.

INTRODUCTION

An important phase in the study of discharge phenomena between contacts is that involving the break of an inductive circuit. A typical switching circuit in its simplest form consists of a battery in series with a coil (electro-magnet), a cable or lead and a pair of contacts. Coils now in use may have inductances of the order of tens of henries and may store as much energy as 10^6 ergs. On break of the circuit an appreciable portion of this energy may be dissipated between the contacts through a steady arc, a series of interrupted arcs, a glow discharge or any of their combinations. In most cases, the energies involved are too high to provide satisfactory contact life from the standpoint of electrical erosion.

The discharge transients obtained are usually complex in nature.¹ A close examination of these transients reveals a great deal of rather curious effects that have not been previously considered in detail. This is a presentation of a recent study of the break transient with the primary objective of furnishing some explanation of the more pertinent phenomena involved in terms of the basic concepts of surface emission and gas conduction.

NOTATION

a	Arc radius or equivalent characteristic length of cross section
c	Local capacitance at the contacts
e	Electron charge
i_a	Current density in the arc
i_{th}	Thermionic emission current density
i_{ng}	Normal glow current density
i_{ag}	Abnormal glow current density
$(i_{g\text{Limit}})$	Limiting glow current density preceding glow-arc Limit transition
k	Boltzman constant
l	Local inductance at the contacts
m	Mass of contact metal atom
n	Number of consecutive arcs in <i>one</i> breakdown
r	Resistance of the local contact circuitry
s	Separation between the contacts
t	Time
t_{ch}	Charging time between breakdowns
t_{dei}	Deionization time following an arc
t_g	Glow duration
u_s	Velocity of contact separation
u_{ch}	Charging velocity defined as s/t_{ch}
u_{at}	Velocity of the metal atoms
v	Arc voltage
$\bar{v}_n$	Residual voltage at the contacts following a breakdown of n -consecutive arcs
z	Impedance $(l/c)^{1/2}$
A	Constant in the thermionic equation
A_a	Area of arc spot
C	Circuit capacitance
E	Battery voltage
F	Field strength
I	Current
I_g	Current in a glow discharge
I_m	Minimum arcing current
I_o	Initial closed circuit current
L	Circuit inductance
R	Circuit resistance
T	Absolute temperature
T_b	Absolute boiling temperature

T_o	Absolute initial temperature
V	Voltage
V_{ai}	Arc initiation voltage
V_{gi}	Glow initiation voltage
V_g	Voltage drop across the contacts with normal glow
α	Thermal diffusivity
φ	Work function
ω	Angular frequency $(lc)^{-1/2}$

GENERAL

A typical circuit consisting of a battery, a coil of an electro-magnet, a cable or lead and a pair of contacts is shown in Fig. 1(a). Due to the usual magnetic core of the coil, this circuit presents some unnecessary complications in making interpretations of the observed contact phenomena. Since our main objective is an understanding of the basic phenomena occurring between the contacts, it appeared justifiable to restrict our work to circuits and circuit elements that lend themselves to simple treatment. Figure 1(b) shows the circuit used in most of this work. All coils used have air cores.

When the contacts are closed, a steady state current $I_o = E/R$ is established in the circuit. At the first physical separation between the contacts, the circuit current will charge the capacitance C causing a voltage rise at the contacts at an initial rate of I_o/C . In the meantime, the separation between the contacts will increase. The first breakdown will occur when the voltage across the contacts first reaches or exceeds the arc initiation voltage corresponding to the separation attained, the atmosphere involved and the *contact surface condition*. Fig. 2 represents diagrammatically the occurrence of the first discharge. *abc* is the arc initiation voltage versus separation line for a "normal" contact.² The

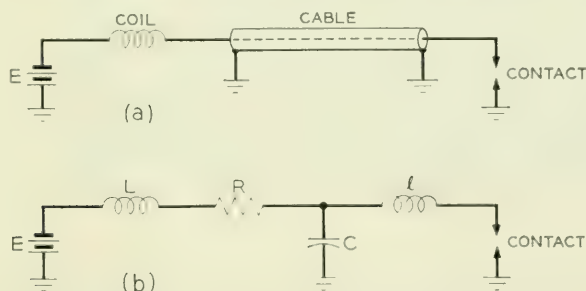


Fig. 1 — (a) Typical relay circuit in practice. (b) Linear circuit used in this study.

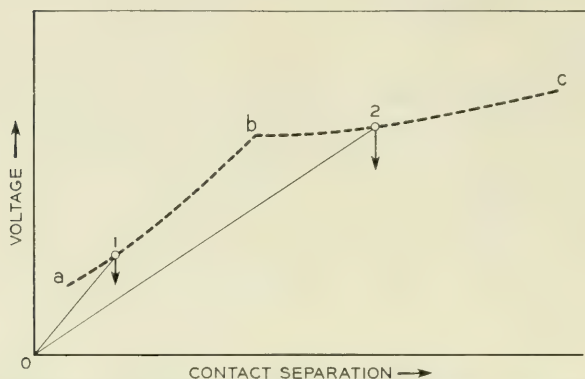


Fig. 2 — Initiation of the first arc between contacts on break of an inductive circuit.

portion bc corresponds to the sparking potentials in the atmosphere. ab corresponds to the range of small separations, of the order of or less than the mean free path of an electron in the atmosphere, where the arc is initiated by field emission through the influence of surface contaminations or films. As was shown in Reference 2, when the cathode surface was carefully cleaned, the constant field line was not obtained and the arc was initiated at the minimum sparking potential of the atmosphere. It occurred on the sides of the contacts along a path much longer than the minimum separation between the contacts.*

Lines 0-1 and 0-2 represent the voltage rise at the contact with small and large shunt capacities. Points 1 and 2 are the respective first discharge points. In the first case, the arc is initiated at a smaller separation and higher field strength without direct influence of the atmosphere. In the second case the arc is initiated at a lower field strength at the spark potential of the atmosphere.†

The first arc established may or may not be maintained depending on conditions that are discussed in the next Section. When an arc is inter-

* With Pd contacts a gross field of 20×10^6 volts/cm was reached between clean contacts without initiating an arc along the shortest gap. According to the Fowler-Nordheim equation a field of about 50×10^6 volts/cm is required to give the necessary initiatory electrons. It is possible, however, that before such a high field is attained a metal bridge is pulled electrostatically³ to short the gap. The electrostatic stress is roughly given by $0.5 \times 10^{-12} F^2$ Kg/cm² where F is the field strength in volts/cm. At $F = 50 \times 10^6$ volts/cm, the stress is 1250 Kg/cm² which may exceed the yield stress for the contact metal.

† The first arc may be initiated at an appreciably lower voltage than predicted by the above static consideration. The first break at the contacts usually follows the explosion of molten bridge drawn between the contacts. Thermionic emission can then furnish the initiatory electrons of the arc. This is only possible, however, if the voltage across the contacts exceeds the ionization potential of the metal atoms before excessive cooling of the cathode has occurred.

rupted, it is followed by a recharging process to a new arc initiation voltage when a second arc is initiated. Under certain conditions, the second arc may be initiated at a lower voltage than the first arc due to residual effects of the first arc which may alter the conditions in the gap. This effect is discussed later.

A transient on break with a series of interrupted arcs is shown in Fig. 3. The first arc was initiated at 230 volts and a gross field of 2.5×10^6 volts/cm. All the following arcs were initiated at the spark breakdown potentials in air corresponding to the separations involved. Fig. 4 shows a transient where the arc was sustained with occasional interruptions.

In addition to arcing, one may obtain glow discharge. Fig. 5 shows a transient where glow discharge predominates. Glow initiation and glow-arc transitions are discussed in a later Section.

Fig. 6 shows the methods used for current and voltage measurements. As indicated, direct voltage measurements at the contacts were avoided to eliminate the unnecessary complications of the measuring circuit.

INTERRUPTED ARCS

Conditions for Obtaining Interrupted Arcs

A breakdown from a voltage V_{ai} into an arc corresponds to a rapid voltage drop at the contacts from V_{ai} to the arc voltage v . For most prac-

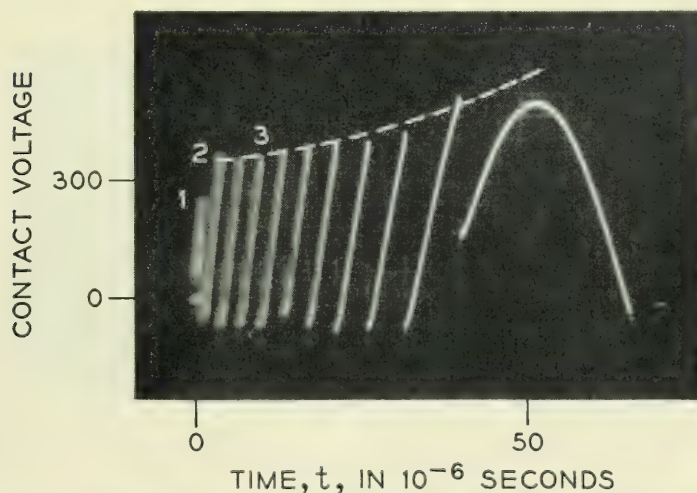


Fig. 3 — Typical contact voltage transient on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, $L = 0.2$ henry, $R = 950$ ohms and $C = 510 \times 10^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

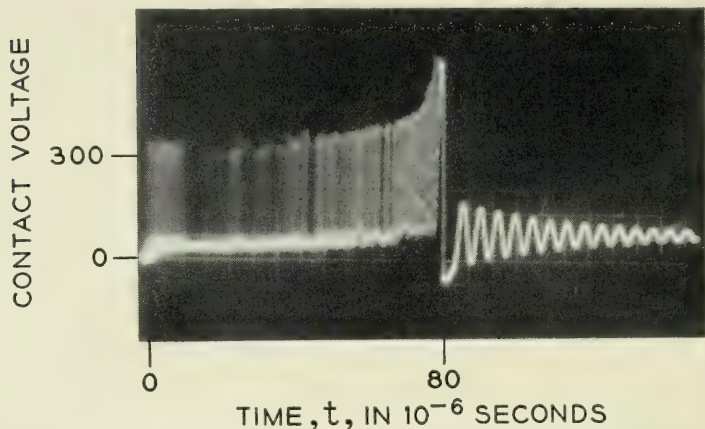


Fig. 4 — Contact voltage transient with sustained arc on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, $L = 0.025$ henry, $R = 115$ ohms, $C = 20 \times 10^{-12}$ farad. Velocity of contact separation 40 cms/sec.

tical purposes one may neglect the voltage drop time which is the initiative period of the arc. For the circuit in Fig. 1b, the current through the arc is the summation of the main circuit current and the transient current from the l - c circuit. The transient current is $(V_{ai} - v)(\frac{C}{l})^{1/2} \sin t/(lc)^{1/2}$. Fig. 7, (a) and (b), represent diagrammatically the voltage and current transients for lumped and distributed circuits. In both cases the arc is

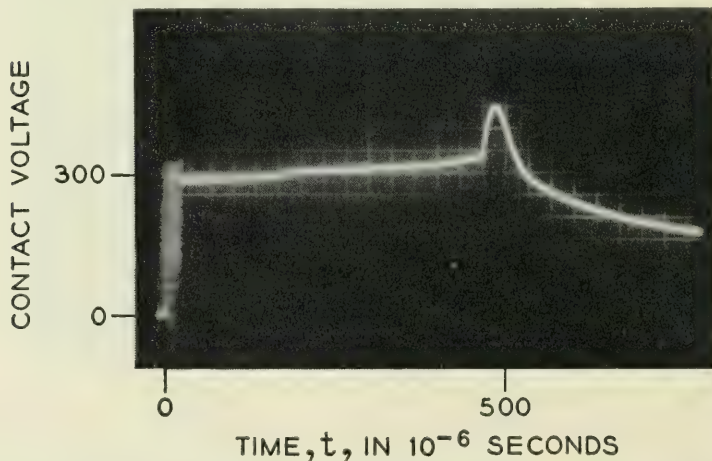


Fig. 5 — Contact voltage transient with glow discharge on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, 700 ohms relay coil and $C = 200 \times 10^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

terminated when the current drops to the minimum arcing current I_m . It is evident that the condition for obtaining an interrupted arc is:

$$I_0 - (V_{ai} - v) \left(\frac{c}{l} \right)^{1/2} < I_m \tag{1}$$

It may be pointed out that surface contamination, such as organic activation, tends to decrease both I_m and V_{ai} ^{2, 4}. According to equation 1, one may conclude that contact surface contaminations usually tend to cause a transition from an interrupted arc transient to a steady arc transient. The latter is usually associated with appreciably higher energy dissipation between the contacts and much lower contact life due to erosion.

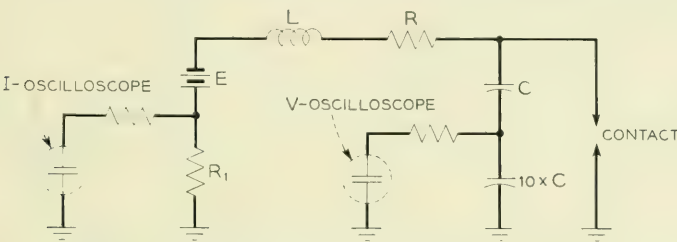


Fig. 6 — Voltage and current measuring circuit.

Residual Voltage Following an Interrupted Arc*

At the interruption of the first arc the voltage at the contact is v , the arc voltage, and the voltage at the capacitor C , Figure 1b, is $\bar{v}_1$ which is usually negative. If the local contact circuit is non-dissipative, the residual voltage is $\bar{v}_1 = 2v - V_{ai}$. For a dissipative circuit with a resistance r corresponding to the frequencies involved:

$$\bar{v}_1 = v - (V_{ai} - v)e^{-(\pi/2) \cdot (r/lz)} \tag{2}^\dagger$$

for an oscillating circuit, as is usually the case, where $z = (L/C)^{1/2}$. The capacitor C at $\bar{v}_1$ will then recharge the local contact capacity c , $c \ll C$, through the inductance l . If the voltage attained at the contacts is sufficient and the conditions in the gap and at the contact surface are favorable, a reversed arc may be re-initiated, as previously discussed. This process may repeat several times and the residual voltage $\bar{v}_n$ will change sign and decrease progressively. At the end of n arcs, it can be shown that the residual voltage $\bar{v}_n$ is given by:

* The term "recovery" has also been used in the literature.
† Equation 2 and 3 are valid only for small values of r/z . These are approximations of the more general expression given by Germer.¹⁴

$$(-1)^n \bar{v}_n = v + (V_{ai} - v)e^{-(\pi/2) \cdot (r/z) \cdot n} - 2v \sum_{n=0}^{n=n-1} e^{-(\pi/2) \cdot (r/z) \cdot n} \quad (3)$$

This equation indicates that $\bar{v}_n$ is negative for odd numbers of arcs and positive for even numbers of arcs.* If r/z is neglected, Equation 3 is reduced to

$$(-1)^n \bar{v}_n = V_{ai} - 2vn \quad (3a)$$

For $V_{ai} = 300$ volts and $v = 14$ volts, the residual voltages following the first four arcs are respectively -272 , $+244$, -216 and $+188$. These

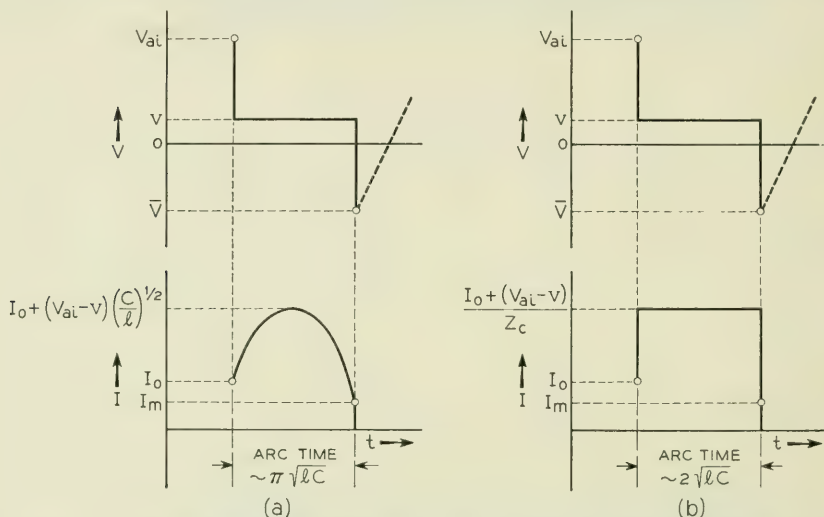


Fig. 7 — Mechanism of interruption of an arc. (a) Lumped circuit elements (b) Distributed elements.

values are numerically higher than measurements due to neglecting the term r/z . For the circuits used in our experiments r/z ranged between 0.1 and 0.5 and as many as 4 or 5 consecutive arcs have been obtained in one breakdown. Figure 8 shows a transient with both positive and negative residual voltages corresponding to even and odd number of arcs respectively.†

* Except when $\bar{v}_n$ is not too much higher than the arc voltage v .

† The following alternative explanation for the occurrence of high positive residual voltage was considered: the first arc may be extinguished by the formation of a metal bridge due to the arc². This may occur before the capacitor C has attained a negative voltage. This possibility, however, was eliminated. From the measured residual voltages the energies in the arcs were calculated. The heights of the bridges produced were computed (reference 2) and were found to be too small compared with the contact separations.

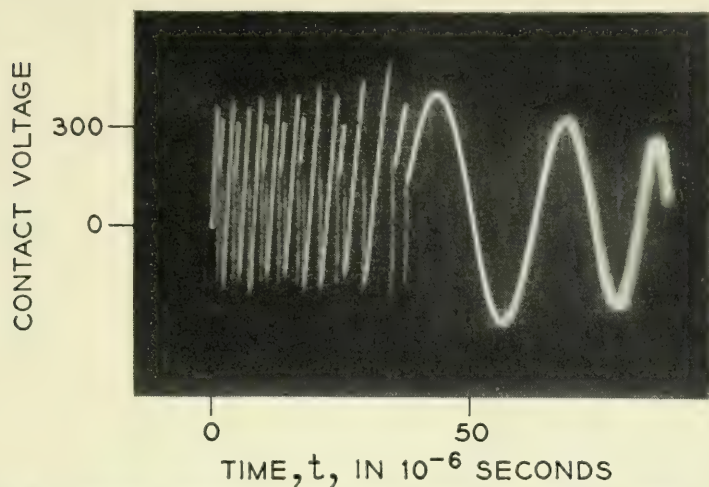


Fig. 8 — Contact voltage transient with interrupted arcs on break of an inductive circuit. Pd contacts in atmospheric air, $E = 50$ volts, $L = 0.010$ henry, $R = 40$ ohms and $C = 900 \times 10^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

Initiation of Reversed Arcs in One Discharge

In one breakdown from a voltage V_{ai} it is commonly observed that a succession of reversed arcs may be obtained. It was shown in equation 3 that the residual condenser voltage $\bar{v}_n$ progressively decreases, numerically, with the number of arcs n . Following the interruption of the first arc, the condenser voltage is $-|\bar{v}_1|$ and the contact voltage is $+v$, the arc voltage. The capacity C will then recharge the local capacitance at the contact through a small lead inductance l . If the circuit resistance is neglected, the maximum voltage the contact will acquire is $-(2|\bar{v}_1| + v)$. If this equals or exceeds the original arc initiation voltage V_{ai} , a second arc is obtained. For illustration, consider a breakdown initiated at $V_{ai} = 300$ volts and $v = 14$ volts. From Equation 3, $\bar{v}_n$ was calculated for the first four arcs at $r/z = 0.0$ and 0.2 . The corresponding maximum contact voltages acquired after each arc were also calculated and the results are given in Table I. For $r/z = 0$, column 3, one may obtain, according to this simple circuit consideration, more than 4 arcs, actually 5. For $r/z = 0.2$, which is a reasonable practical value, only 2 arcs may be obtained, column 5, since following the second arc the maximum voltage attained at the contacts is only 256 volts which is less than the initial arc initiation voltage.

It is possible in some cases, however, to obtain a few more arcs than

TABLE I — INITIATION OF REVERSED ARCS BY OVERCHARGING OF CONTACT CAPACITANCE
(Calculated)

$\frac{r}{z} = 0$			$\frac{r}{z} = 0.2$	
(1) Arc No.	(2) v_n	(3) Max. Cont. Voltage	(4) v_n	(5) Max. Cont. Voltage
1	-272	-558	-195	-404
2	+244	+502	+121	+256
3	-216	-446	-60	-134
4	+188	+390	+22	+58

$V_{as} = 300$ volts, $v = 14$ volts.

predicted above. These additional arcs have appeared to be initiated at lower voltages than the first arc. This is undoubtedly due to the residual surface and gap effects of the previous arc.* These are discussed in the following section.

Arc Initiation Under Dynamic Conditions — Introduction

In Reference 2 measurements have been presented of the arc initiation voltage between contacts at different separations and surface conditions. These tests are "static" in the sense of allowing enough time to elapse between two arcs to obtain a complete reconditioning of the contact surfaces and gap. With successive arcing, as obtained on break of an inductive circuit or during one breakdown, it was observed that the arc may be initiated at appreciably lower voltages compared with static test results.

One arc may enhance the initiation of a shortly following arc possibly through the effects of: residual ions in the gap or on a cathode surface film, residual metal atoms in the gap and residual thermionic emission. Exactly how each of these effects can enhance the initiation of the arc can be determined only after an understanding of the mechanisms of initiation of the first arc, its maintenance and its termination. It is in order at this point to present a sketchy outline of some plausible mechanisms which are largely of speculative nature. This discussion is also limited to short arcs initiated and maintained with no direct influence of the surrounding atmosphere.

* The additional arcs observed may be partially accounted for by a consideration of the actual value of the arc terminating current which was taken as zero in the above calculations.

a. Arc Initiation

(1) The first initiatory electrons are produced by field emission. The necessary field strength is largely dependent on cathode surface conditions. It is highest for perfectly clean cathode surfaces and appreciably lower in the presence of cathode surface films.^{2, 4, 6} This is probably due to lower work functions or due to the presence of positive ions on a cathode film causing local field intensification.⁷ (2) The field emission electrons will travel to the anode where, to qualify for setting the second step in arc initiation, should be able to produce, through evaporation, some anode metal atoms* or possibly atoms of an adsorbed gas or a surface film. (3) The potential drop across the contacts should exceed the ionizing potential of the evaporated atoms to allow ionization by electron collision. (4) Ions produced, on approaching the cathode, will cause local fields high enough to produce electron avalanches. (5) the above processes will rapidly multiply leading to the establishment of an arc.

b. The Established Arc

One main characteristic of the short arc is its very high cathode current density.† This high emission rate indicates that the short arc is not only initiated *but also maintained by field emission.*‡§ Since the total voltage drop across the arc is only of the order of 10 volts, the cathode drop thickness should be very small compared to the total arc length. The cathode drop is followed by the arc column or plasma which is a high conduction medium with equal electron and ion densities, a small potential drop and a relatively high neutral atom density. To maintain the arc: (1) enough metal atoms should be produced to maintain the necessary ionization medium, (2) ions lost by collection at the cathode, by recombination and by lateral diffusion should be replaced by an

* The arc may also be initiated without the assistance of the anode atoms or ions⁸. The field emission current density at the cathode in this case, was found to reach a critical value before the arc is initiated. It is thought⁹ that at this current density the emission spot can attain its melting point through resistive heating. The cathode in this case will furnish the necessary metal atoms for the subsequent steps of arc initiation.

† Recent measurements by the author obtained from arc tracks on Pd contacts produced by short duration *constant* current arcs indicated current densities as high as 50×10^6 amp/cm².

‡ Paper by P. Kislink to be published in the *Journal of Applied Physics*.

§ Recent analytic considerations, to be published by the author, indicate that in such arcs the current density should be dependent on the work function of the cathode material as well as on the product "pressure $\times$ separation" in the arc. For instance, for work functions of 2 and 5 volts, our calculations show that the minimum current densities are, respectively, 5×10^5 and 1.4×10^7 amp/cm².

equal number of ions obtained by electron-atom collision in the arc column.

c. Arc Termination

In general, the arc may be terminated by disturbing one or more of the steady state conditions discussed above. For instance, if the potential across the contacts is decreased to or below the ionization potential of the metal atoms, the necessary ionization process will stop and a deficiency of ions in the arc will result. The negative space charge will immediately upset the arc potential distribution interrupting the high electron emission, etc. The arc is also interrupted when the current drops to the minimum arcing current value. This is a well established experimental characteristic of the arc which has yet to be explained in terms of the more basic concepts. It is thought, however, that a decreasing arc current decreases the pressure and the atom density in the arc column. It is possible that when a limiting current is reached the ionization rate becomes too small to maintain the condition of equal space charges in the arc column. One should expect, accordingly, that providing the contact surfaces* with a film of low evaporation energy should furnish a more adequate supply of atoms to the arc which may then be maintained at lower currents. This is in accordance with observations obtained for active contacts.⁴

Arc Initiation Under Dynamic Conditions; Observations on Break

It appeared of interest to examine the relations between arc initiation voltage and contact separation during the break transient and compare them with measurements made under static conditions.² In Fig. 3, the increase in arc initiation voltage with separation is in accordance with the static relation shown as a broken line. During the period 2-3, the breakdowns occurred along longer paths than the minimum contact separation and at the minimum value of the sparking potential. By measurement $t_3 = 20 \times 10^{-6}$ sec, $s_3 = 8 \times 10^{-4}$ cm and $ps = 0.61$ mm Hg $\times$ cm. This is roughly the ps value at the minimum sparking potential in air.¹⁰

By gradually decreasing the charging times of the transient, by adjusting circuit parameters, it was observed that a point was generally reached when a portion of the breakdowns was initiated at voltages well below the corresponding static initiation voltages. Fig. 9 illustrates this

* The necessary atoms may be obtained from either electrodes or both. Arc transfer observations generally indicate signs of evaporation from both electrodes.

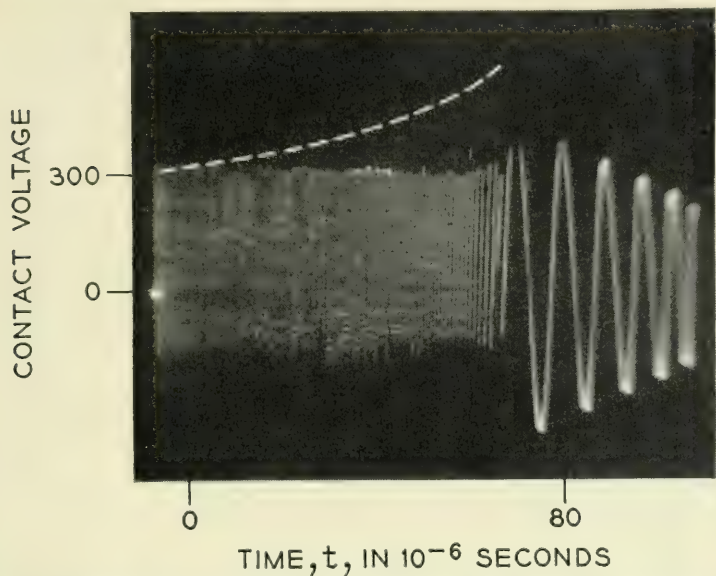


Fig. 9 — Lowering of arc initiation voltage under dynamic conditions. Transient on break of Pd contacts in atmospheric air. $E = 50$ volts, $L = 0.010$ henry, $R = 40$ ohms and $C = 270 \times 10^{-12}$ farad. Velocity of contact separation = 40 cms/sec.

effect. In contrast to the static line, shown as a broken line, the break-down potential shows little change with separation for a major part of the transient. Towards the end, it shows a gradual increase which in this particular case fails to reach the static line. Figure 10(b) is a plot of the ratio $(V_{ai})_{dyn}/(V_{ai})_{stat}$ versus time along the transient.

This phenomenon is attributed to residual effects in the contact gap or on the contact surfaces. In this section, are discussed the possibilities of the presence of residual ions, residual atoms and residual thermionic emission.

a. Deionization Time

This is determined by calculating the transit time of an ion across the contact gap under the applied field corresponding to the charging of the contact capacitance. For simplification, the initial motion of the ions and the initial field are neglected, the voltage rise is approximated by $V/V_{ai} = t/t_{ch}$ and the field is taken as V/s .

$$t_{deio} = \left(\frac{6m}{e} \cdot \frac{s^2 t_{ch}}{V_{ai}} \right)^{1/3} \tag{4}$$

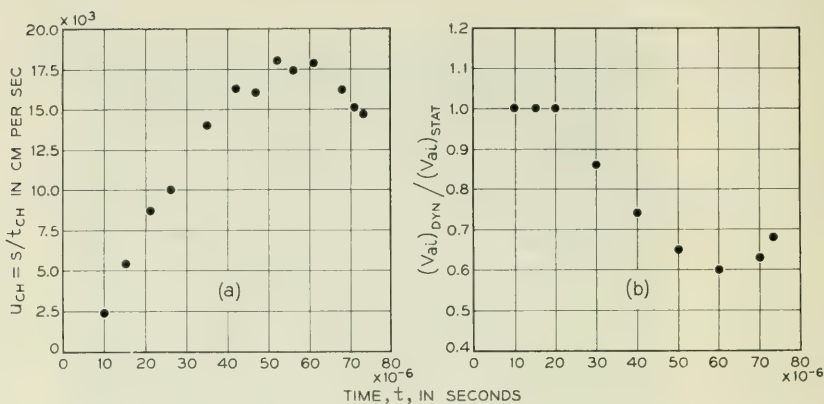


Fig. 10 — Lowering of arc initiation voltage under dynamic conditions.

Defining a charging velocity $s/t_{ch} = u_{ch}$ and a deionization velocity $s/t_{deio} = u_{deio}$ and substituting in equation 4 gives

$$u_{deio} = \left(\frac{eV_{ai}}{6m} \cdot u_{ch} \right)^{1/3} \quad (4a)$$

Following an arc, the contact voltage increases until a new breakdown occurs at V_{ai} . At this instant residual ions from the previous arc could be present in the gap only if $u_{ch} > u_{deio}$, or if

$$u_{ch} > \left(\frac{eV_{ai}}{6m} \right)^{1/2} \quad (5)^*$$

This is a convenient expression to apply to our measurements, Fig. 9. For any breakdown point on the transient V_{ai} is measured and u_{ch} is calculated from the corresponding circuit current, capacity C and contact separation. For illustration, for Pd contacts and $V_{ai} = 300$ volts, equation 5 shows that for the presence of residual ions, the charging velocity u_{ch} must be greater than 10^6 cms/sec. For $I = 0.3$ amp. and $C = 10^9$ farad, $t_{ch} = V_{ai}C/I = 10^{-6}$ sec and for the presence of residual ions the separation between the contacts must be greater than 1.0 cm. This separation is much greater than most separations involved in our field of study. In Fig. 10(b) are plotted the values of u_{ch} during the transient. u_{ch} reaches a maximum of about 1.8×10^4 cms/sec. This maximum occurs because u_{ch} is proportional to sI which is a product of two monotonic functions one increasing and the other decreasing. It is of interest to note that the decrease in u_{ch} caused an increase in the ratio $(V_{ai})_{dyn}/(V_{ai})_{stat}$.

* Deionization by recombination and lateral diffusion were neglected.

From a group of transients similar to Fig. 9, obtained at different conditions, the plot in Fig. 11 was made. It indicates that in general, the ratio $(V_{ai})_{dyn} / (V_{ai})_{stat}$ starts decreasing at about $u_{ch} = 2 \times 10^3$ cms/sec and at 2×10^4 the arc initiation voltage is only 50 per cent of the corresponding static value. As shown in the figure a deionizing velocity of 10^6 cms/sec is just about two orders of magnitude too high to account for this phenomenon. It should be added, however, that while all the ions have cleared the gap, it has been proposed⁷ that the life time of an ion on a surface film can be long enough to enhance the initiation of the next arc. If this mechanism is accepted, our data would indicate that the life time of the ions was only of the order of 10^{-7} second.

b. Residual Atoms

After an arc, the contact gap contains some metal atoms evaporated from the electrodes by the arc. These atoms will clear the gap by traveling to and condensing on the electrodes and by lateral diffusion. A crude approximation is given here of the time of recollection of the atoms on the electrodes based on their initial momentum.

One may visualize the arc spot on an electrode to have a temperature distribution extending from submelting temperatures to a range of boiling temperatures, corresponding to the arc pressures. The lowest temperature is probably the normal boiling temperature of the contact metal. At the termination of the arc, the metal atoms produced at the lowest boiling temperature are the slowest and last to recondense on the

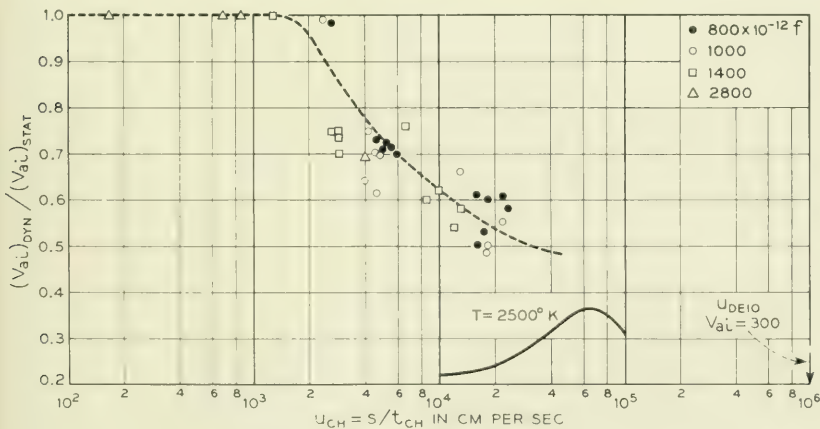


Fig. 11 — Apparent relation between arc initiation voltage and velocity of charging. $E = 50$ volts, $L = 0.010$ henry, $R = 40$ ohms and C as indicated for Pd contacts in atmospheric air.

opposite electrode. An estimate of their velocity may be obtained by assuming thermal equilibrium to have preceded the arc extinction and by using the Maxwellian velocity distribution. The most probable velocity of the metal atoms at the boiling temperature T_b is:

$$u_{at} = \left(\frac{2kT_b}{m} \right)^{1/2} \quad (6)$$

Due to subsequent collisions of the atoms, the velocities thus obtained are probably too high. For Pd at $T_b = 2500^\circ\text{K}$, $u_{at} = 6.4 \times 10^4$ cms/sec. In Fig. 11 is plotted a portion of the velocity distribution at the above conditions. It appears that residual atoms can still be present in the gap at the initiation of the next arc. If it is assumed that the presence of Pd atoms in the gap is alone responsible for the lowering of the arc initiation voltage, one may conclude that the sparking potential in Pd vapor is lower than in air. No evidence, however, is available to support this. On the other hand, at least for contacts with gaps short enough to exclude the surrounding atmosphere, or for vacuum contacts in general, it is quite probable that the presence of metal atoms in the gap could enhance arc initiation. This, as pointed out previously, is because the arc cannot be initiated until atoms from the electrode surfaces are evaporated, by electron bombardment or otherwise, to be subsequently ionized.

c. Cooling Time of The Arc Spot, Maintenance of Thermionic Emission

At the interruption of the first arc, the arc spot initially at the boiling temperature of the metal, will start cooling mainly by conduction to the bulk of the surrounding metal. For a certain period, however, it will remain at temperatures high enough to furnish enough thermionically emitted electrons that may enhance the initiation of the following arc. Assuming the arc spot to be a hemisphere of radius " a " initially at a temperature T_b while the rest of the metal is at T_o , the temperature T at the center of the hemisphere is given by ¹¹:

$$(T - T_o)/(T_b - T_o) = \frac{4}{\pi^{1/2}} \int_0^{a/2(\alpha t)^{1/2}} z^2 e^{-z^2} dz \quad (7)$$

Numerically, for $T_b = 2500^\circ\text{K}$ and $T_o = 300^\circ\text{K}$, T drops to 2400°K and to 1600°K at $a/2(\alpha t)^{1/2} = 2.0$ and 1.2 , respectively. It is evident that the cooling time is proportional to the area of the arc spot. If the current at which the arc is terminated is I_m and the arc current density is i_a , the area of the arc spot is $A_a = I_m/i_a$ and $a = (I_m/\pi i_a)^{1/2}$. For

$i_a = 10^7$ amp/cm²⁷, and $I_m = 0.5$ amp one gets: $T = 2400^\circ\text{K}$ at $t = 4 \times 10^{-9}$ sec and $T = 1600^\circ\text{K}$ at $t = 1.1 \times 10^{-8}$ sec for Pd.

The corresponding thermionic emission is obtained from

$$i_{th} = AT^2 e^{-\frac{e\phi}{kT}}$$

with $A = 60$ amp cm⁻² deg.⁻² and $\phi = 4.99$ volts for Pd¹². At the termination of the arc, $t = 0$, $i_{th} = 0.048$ amp/sec², at $t = 4 \times 10^{-9}$ sec, $i_{th} = 0.032$ amp/cm² and at $t = 1.1 \times 10^{-8}$ sec, $i_{th} = 6 \times 10^{-8}$ amp/cm². The respective rates of electron emission from the arc spot are 1.5×10^{10} , 1.0×10^{10} and 1.9×10^4 electrons/sec. *This indicates that the initiating electrons may be furnished by thermionic emission if the charging time following the first arc is of the order of or less than about 5×10^{-9} sec.* This time is more than an order of magnitude too small compared to the charging times involved in the data of this section. One may, therefore, exclude the thermionic emission as an explanation for the low arc initiation voltages obtained.

The initiation of reversed arcs, however, may be enhanced by thermionic emission from the previous arc spots since the recharging times involved, $\pi(lc)^{1/2}$, are usually very small. l and c are usually of the orders of 10^{-7} henry and 10^{-11} farad and the charging time is of the order 10^{-9} sec.

ESTABLISHMENT OF GLOW DISCHARGE AND TRANSITION INTO AN ARC

For the circuit in Fig. 1(b), it was observed that on break of the contact, glow discharge was observed under certain circuit and contact surface conditions. An obvious requirement was that the voltage across the contacts should exceed the glow discharge voltage of the contact in the surrounding atmosphere. This requirement alone, however, was not sufficient as in some cases no glow could be detected, in others glow was established and maintained and in other instances glow was followed by a transition into an arc. In this section is presented an experimental study of the conditions that determine the nature of the discharge.

Cathode Current Density in Static Normal Glow

First, measurements were made of the cathode current density in a static normal glow. This was done for palladium and gold contacts in dry atmospheric air at 25°C . In each case the cathode was the flat end of a cylinder and the anode was a larger parallel flat surface of the same material as the cathode. The circuit in Fig. 12 was used. The contacts were

cleaned by filing then washing with methyl alcohol and distilled water. The contacts were slowly brought together until glow discharge was established. Before measurements were made the circuit current was increased to allow the glow to cover the entire cathode flat surface as well as a portion of its cylindrical surface. By allowing the contacts to glow for about 20 minutes, the occasional arcing first observed was eliminated and a steady glow was established. The cathode was observed under a microscope and the current was adjusted to obtain a glow just covering the flat cathode area. From the measured current and the cathode area, the cathode current density was determined. The results are given in Table II.

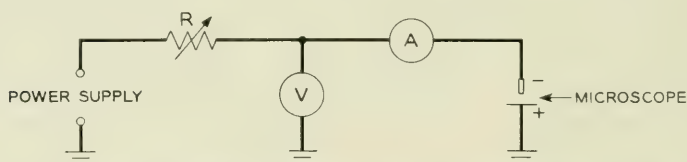


Fig. 12 — Circuit for measurement of cathode current density in normal glow discharge at static conditions.

Observations on Glow Maintenance and Glow-arc Transitions

The simplified circuit in Fig. 13 was used. The contact cathode was the flat end of a cylinder. The cylindrical portion was tightly fitted into a block of an insulating material allowing an exposure of the flat end and a cylindrical area less than 10 per cent of the flat area. The anode was a parallel plain surface of the same material.

To avoid the unnecessary complications of a measuring circuit connected to the contacts, the plates of a cathode ray oscilloscope were, instead, connected across a capacitor, 10 times C , in series with the circuit capacitor C . From the transients obtained, it was possible to identify glow discharge, steady arcs and interrupted arcs. Four typical transients are shown in Fig. 14. Transient A shows a case where glow discharge was established and maintained for the entire half period of the circuit. In transient B glow was not detected and, instead, interrupted arcs occupied the entire half period. In transient C, glow discharge was maintained for a short duration 1-2 followed by interrupted arcing, 2-3. At point 3 the circuit current was high enough to maintain an arc and a steady arc was obtained, 3-4. Transient D is similar to B where glow discharge was undetectable. The multiple discharge in D, however, lead to the steady arc 2-3.

Before presenting our measurements and discussion, a review is given

TABLE II — CATHODE CURRENT DENSITY IN STEADY NORMAL GLOW
IN DRY ATMOSPHERIC AIR AT 25°C

Electrodes	Cathode Diameter, cm.	Glow Current, amp.	Cathode Current Density, amp./cm ²
Pd	0.05	0.010	5.1
	0.10	0.033	4.2
Au	0.05	0.017	8.6
Ag ^a	—	—	9

* Measurement by F. E. Haworth.¹³

here of the process of the initiation of the steady arc which was explained in detail in reference.⁵ For the inductive circuit in Fig. 13, when the proper contact separation is reached, a first breakdown will occur discharging the local capacitance at the contacts. This is followed by recharging from C through L and a second breakdown. This will repeat while the circuit current will increase in a discontinuous fashion. If it reaches the minimum arcing current of the contact, a steady arc is established, otherwise, the transient will be made up entirely of local multiple discharges. Figures 14D and B are the main condenser voltage transients corresponding to the above two cases respectively.

The interrupted arcs, or multiple discharges, and the steady arc constitute the two processes of conduction that are commonly obtained when the voltages involved are below the spark breakdown potential of the surrounding atmosphere. In such cases, the arc initiation is dependent on the contact material and its surface condition and is independent of the atmosphere.² If the voltages involved are equal to or greater than the minimum sparking potential of the atmosphere, the initiation of a breakdown is primarily dependent on the atmosphere. This breakdown, however, may in addition lead to a glow discharge as discussed above. This immediately raises the question as to whether breakdowns leading to an arc and breakdowns leading to a glow discharge are initiated at the same potentials. For this purpose the following experiment was performed.

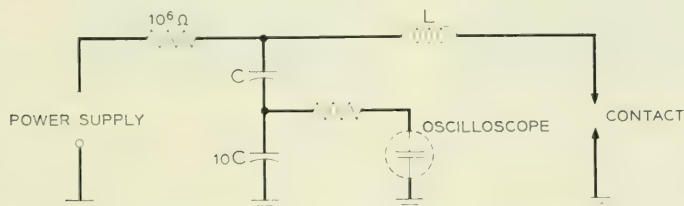


Fig. 13 Simplified circuit for the study of glow-arc transitions.

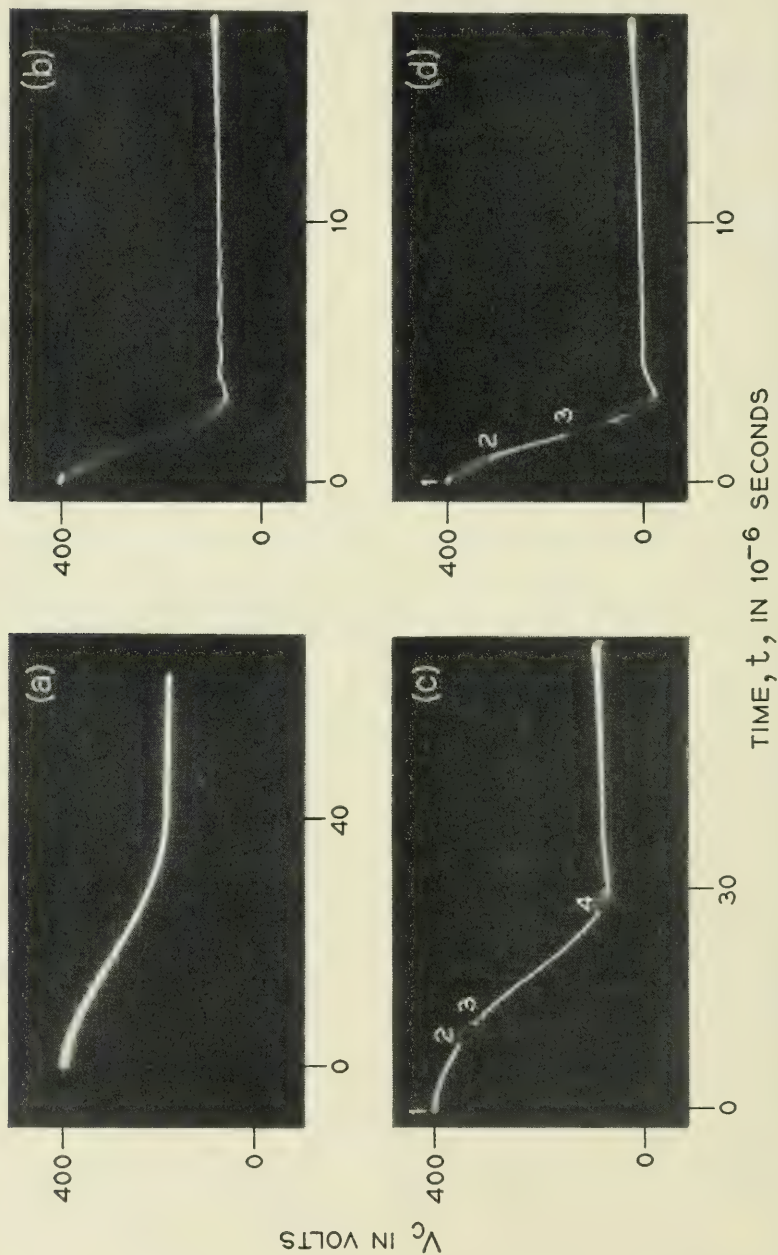


Fig. 14 — Voltage transients at circuit condenser. A: all glow. B: no glow. C: glow-inter-rupted arcs — steady arc. D: no glow, interrupted arcs — steady arc — interrupted arcs.

Initiation Voltage of Glow Discharge

The cantilever bar setup previously used for similar measurements of arc initiation voltage as function of separation² was used here. By varying the separation the corresponding glow initiation voltage was measured. For each separation a measurement was also made of the arc initiation voltage. The results are given in Table III. The results indicate that both arc and glow are initiated at the same voltage for the same separation. One may, therefore, conclude that *at least the first few steps involved in the process of the breakdown are the same whether they lead to a glow discharge or to an arc*. In many cases, it was observed that the arc was preceded by a period of glow discharge. This was not found, however, to be the general case as discussed in the following section.

TABLE III — GLOW AND ARC INITIATION VOLTAGES AS FUNCTIONS OF CONTACT SEPARATION FOR Pd CONTACTS IN DRY ATMOSPHERIC AIR AT 25°C.

S: 10 ⁻⁴ cm.	1.5	3.0	4.5	6.0	7.5	9.0	10.5	12.0	13.5	15.0	16.5	18.0	19.5
V _{oi} : volts	320	320	340	380	400	420	450	480	500	520	540	560	590
V _{ai} : volts	310	320	340	370	400	420	450	470	490	510	540	570	590

Glow-arc Transition

The experimental setup used is shown in Fig. 13. By systematic variation of the circuit parameters V_o , L and C , a variety of transients was obtained and recorded. Samples of typical cases are shown in Fig. 14. For transient stability and reproducibility, it was found necessary to exercise extreme care in securing good contact surface cleanliness and in maintaining it during the experiment. The presence of organic vapors, humidity, films of grease or oil, fingerprints, etc., usually led to erratic results. *The general effect was an inclination towards more arcing and less glow discharge*. Only by proper cleaning of the contact surfaces and allowing the contact to arc heavily for about 20 minutes was it possible to obtain fairly reproducible results. Table IV shows a summary of results obtained from one of several sets of experiments performed.

Before stabilization of the transient, it was generally observed that the glow period was first short then gradually increased until it reached a limiting value which it did not exceed. These limiting values are given in column 5 as fractions of the half period $\pi(LC)^{1/2}$. They range from zero, actually glow was not detected with a time resolution of 1 per cent of the

TABLE IV — GLOW-ARC TRANSITION DATA FOR Pd CONTACTS
IN ATMOSPHERIC AIR-CATHODE DIAMETER 0.1 CM.

(1) V_0 volts	(2) C $10^{-12}f$	(3) L $10^{-3}h$	(4) $= \frac{Z}{(L/C)^{1/2}}$ ohms	(5) $I_0/\pi (LC)^{1/2}$	(6) $(V_0 - V_g)/L$ 10^4 amp./sec.	(7) (I_0) max. amp.	(8)† (i_0) max. amp./cm. ²	(9) $(i_0)/i_0$ max.
600	18,000	5	52	0*	6.0	<.018	<2.3	<0.5
500	—	—	—	0	4.0	<.012	<1.6	<0.4
450	—	—	—	0	3.0	<.009	<1.2	<0.3
400	—	—	—	1.0†	2.0	>.19	>23	>4.8
350	—	—	—	1.0	1.0	>.10	>13	>2.6
320	—	—	—	1.0	0.4	>.04	>5	>1.0
600	18,000	8	660	0	3.8	<.015	<2	<0.4
500	—	—	—	0.22	2.5	.19	24	4.8
450	—	—	—	0.44	1.9	.23	29	5.8
400	—	—	—	1.0	1.2	>.15	>19	>3.8
350	—	—	—	1.0	0.6	>.076	>10	>2.0
600	18,000	15	906	0.25	2.0	.23	29	5.8
550	—	—	—	0.30	1.7	.23	29	5.8
500	—	—	—	0.35	1.3	.20	26	5.2
450	—	—	—	1.0	1.0	>.17	>22	>4.4
400	—	—	—	1.0	0.7	>.11	>14	>2.8
600	18,000	20	1050	0.40	1.5	.28	36	7.2
550	—	—	—	0.40	1.2	.23	29	5.8
500	—	—	—	1.0	1.0	>.19	>24	>4.8
450	—	—	—	1.0	0.8	>.14	>18	>3.8

* No glow was detected with a time resolution of 1 per cent of a half period $\pi(LC)^{1/2}$.

† Uninterrupted glow occupied the entire half period.

‡ Obtained by dividing $(I_0)_{\max}$ by the total cathode area.

transient time, to a full transient time. By calculation, the corresponding limiting currents and limiting current densities were obtained, columns 7 and 8 respectively. The ratios of the limiting current densities to the normal glow current density are also given in column 9. They show that at the interruption of the glow discharge the current density was 5 to 7 times the normal glow current density. This indicates a transition from normal glow to abnormal glow before the final transition into an arc. One may, therefore, conclude that if glow discharge is obtained it starts as normal glow which may occupy only a small fraction of the cathode area. By increasing the current the cathode glow area expands at constant current density until it covers the entire cathode area. Further current increase leads to a transition into abnormal glow with higher current densities. Transition of the abnormal glow into an arc occurs when the current density reaches a limiting value. This limiting current density is extremely sensitive to surface contamination and generally

increases with surface cleaning.* For clean Pd contacts in atmospheric air an average limiting current density of 30 amps/cm², or about 6 times the normal glow current density, was obtained. This sudden transition from the low current density glow to the very high current density arc represents a high rate of change in the emission process. With contaminated contacts, this is probably due to the presence of low work function high emission spots on the cathode. These spots may be eliminated by proper cleaning thus allowing glow discharge to be maintained at higher current densities. The observed glow-arc transitions for clean contacts, consistently occurring at about 30 amps/cm² for Pd, *may still be attributed to the formation of a surface film on the cathode through a cathode-atmosphere reaction.*†

Measurements have also indicated that under certain conditions, glow discharge cannot be obtained even at currents much below the limiting currents discussed above. It appears that there is a limiting rate of rise of current with time above which glow discharge cannot be maintained. In Table IV, column 6, the initial rates of current rise are given. In all cases where the rate of current rise was greater than about 3×10^4 amps/sec, lines 1, 2, 3 and 7, no glow was obtained. The experiment was repeated with two other cathode diameters of 0.2 and 0.05 cm. The limiting rates of rise obtained were approximately the same as given above, indicating that the limiting rate of current rise is independent of the cathode area. This seems reasonable since at the beginning of the transient the currents are very small and the emission area is only a very small fraction of the cathode area. No detailed explanation, however, can be furnished at this time as to why such a limit of the rate of current rise does exist. It is obvious, nevertheless, that while the rate of current rise can be increased without limit by manipulating the circuit parameters, the conduction mechanism in the contact gap, will, in general, have its own limitations as determined by the emission processes involved.

ACKNOWLEDGMENT

I am indebted to Miss R. E. Cox for assistance with many of the experiments and calculations reported here.

* With a contaminated cathode surface a transition into an arc may occur during the *normal* glow period well before the current is high enough to allow normal glow to cover the entire cathode surface. This is particularly true with larger cathode areas which are usually hard to clean satisfactorily by the above procedure.

† A recent unpublished study by F. E. Haworth has shown that in the absence of the usual surface contaminants, glow discharge is capable of activating palladium and silver contacts through the formation of surface films. These surface reactions appear to be strongly dependent on the atmosphere.

BIBLIOGRAPHY

1. A. M. Curtis, Contact Phenomena in Telephone Switching Circuits, B. S. T. J., **19**, p. 40, 1940.
2. M. M. Atalla, Arcing of Electrical Contacts in Telephone Switching Circuits — Part II, B. S. T. J., **32**, pp. 1493–1506, Nov., 1953.
3. G. H. Pearson, Phys., **32**, pp. 1493–1506, Rev., **56**, p. 471, 1939.
4. L. H. Germer, Arcing of Electrical Contacts on Closure — Part I, J. Appl. Phys., **22**, p. 955, 1951.
5. M. M. Atalla, Arcing of Electrical Contacts in Telephone Switching Circuits — Part I, B. S. T. J., **32**, p. 1231, 1953.
6. F. E. Haworth, Experiments on the Initiation of Electric Arcs, Phys. Rev., **80**, p. 223, 1950.
7. F. L. Jones, Initiation of Discharges at Electrical Contacts, Proc. Inst. Electrical Engineering I 124, 169, 1953.
8. W. P. Dyke, J. K. Trolan, E. E. Martin, and J. P. Barbour, The Field Emission Initiated Vacuum Arc — I, Phys. Rev., **91**, p. 1043, 1953.
9. W. W. Dolan, W. P. Dyke, and J. K. Trolan, The Field Emission Initiated Vacuum Arc — II, Phys. Rev., **91**, p. 1054, 1953.
10. J. J. Thomson and G. P. Thomson, *Conduction of Electricity Through Gases*, Vol. 2, p. 487.
11. H. S. Carslow, *Introduction to the Mathematical Theory of the Conduction of Heat in Solids*, 2nd Edition, p. 150, 1921.
12. S. Dushman, Rev. Mod. Phys. **12**, p. 381, 1930.
13. F. E. Haworth, Electrode Reactions in Glow Discharge, Jl. Appl. Phys., **22**, p. 606, 1951.
14. L. H. Germer, Erosion of Electrical Contacts on Make, J. Appl. Phys., **20**, pp. 1085–1109, 1949.

Thickness Measurement and Control in the Manufacture of Polyethylene Cable Sheath

By W. T. EPPLER

(Manuscript received October 22, 1953)

The manufacture of multiple sheath for Alpeth and Stalpeth cables requires the application of a sheath of polyethylene over a sheath of corrugated metal which is flooded with a rubber asphaltic compound. For high quality and minimum cost, this outer sheath must be of uniform thickness throughout its length. One of the problems in cable sheath manufacture is to maintain the concentricity and average thickness of the extruded polyethylene sheath to close limits during manufacture. This article reports on: (1) The application of a capacitance sensitive bridge to the measurement of the eccentricity and average thickness of the sheath on cables moving at speeds of 20 to 100 feet per minute; (2) The method of thickness calibration; and (3) The use of the thickness measurements in maintaining the sheath concentricity and average thickness within close limits during the sheathing operation.

HISTORY

In the manufacture of multiple sheath for Alpeth and Stalpeth cables, an outer sheath of polyethylene is applied. It is desirable for high quality and low cost to make this outer sheath of a uniform thickness throughout. The construction of these cables is shown in Fig. 1. In both designs, the outer sheath is polyethylene extruded onto a corrugated metal under-sheath which has been flooded with a rubber asphaltic compound.

The extrusion art had been unable to obtain a high degree of control, primarily because measurements of the thickness could not be obtained until after the sheath was applied to the cable core. Eccentric sheath must have a greater average thickness than concentric sheath, if the thickness of the thin side is not to fall below a required minimum thickness.

The symmetrical design of a typical core tube and die for sheathing is shown in Fig. 2. Concentric set-up of these extrusion tools around the

cable core will not produce concentric extruded sheath. This is caused by an unbalance in the plastic flow in the extruder. The flow makes a ninety degree turn from the extruder cylinder into the die head, and to reach the far side of the die, must flow around the core tube. The flow resistance also varies with changes in the temperature of the plastic and of the extruder screw speed.

The core tube is fixed in position in the extruder head. The die is located around the core tube and can be moved in any direction eccentric to it. Fig. 3 shows a core tube and die mounted in the extruder head and indicates the location of the four die adjusting screws by which movement of the die in relation to the fixed core tube is accomplished. The die must be located at some one eccentric position in relation to the core tube to compensate for the differences in flow resistances in the head.

To set the die for concentric sheath and to adjust for specified thickness the prevailing practice of the cable art of measuring the wall thickness of a sample taken from the lead or finish ends of the sheathed cable was of necessity resorted to because it was the best technique available. The cutting of a ring of sheath and the micrometer gage are shown in Fig. 4. These end samples only approximate sheath conditions because

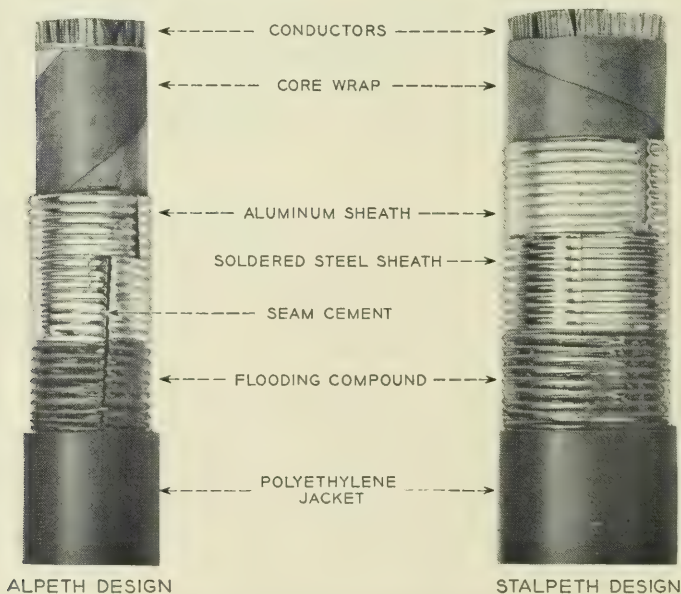


Fig. 1 - (Left) Telephone exchange cable of Alpth design; (right) Stalpth design.

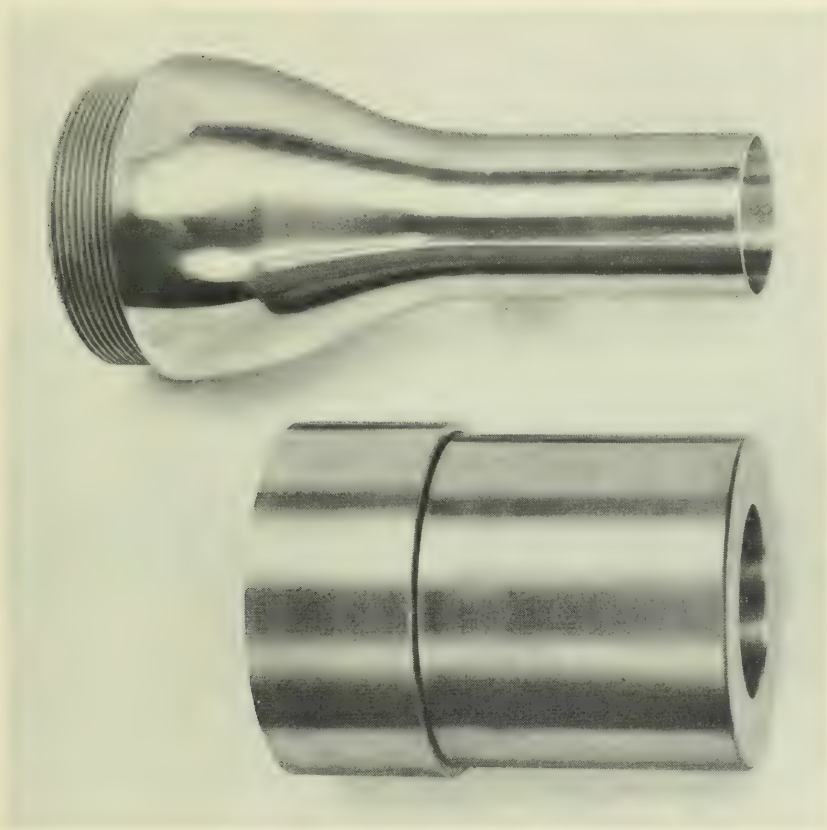


Fig. 2 — Typical core tube and die.

they are only short pieces to represent cables up to a few thousands of feet in length.

Sheath eccentricity is expressed as a percentage and is the difference between the thicknesses, of the thickest and the thinnest sides of a cross section, in relation to the specified wall thickness expressed in mils. Control from end sampling resulted in most cables having eccentricities of 30 per cent to 60 per cent. Also, it was difficult to keep the average thickness to within ± 0.010 inch of the specified average thickness.

The need for a better gaging method than end sampling, led to an investigation of determining the wall thickness in terms of the capacitance that would be formed by the metal undersheath and a probe sliding on the sheath surface.

A test set as shown in Fig. 5 was developed which responds to changes

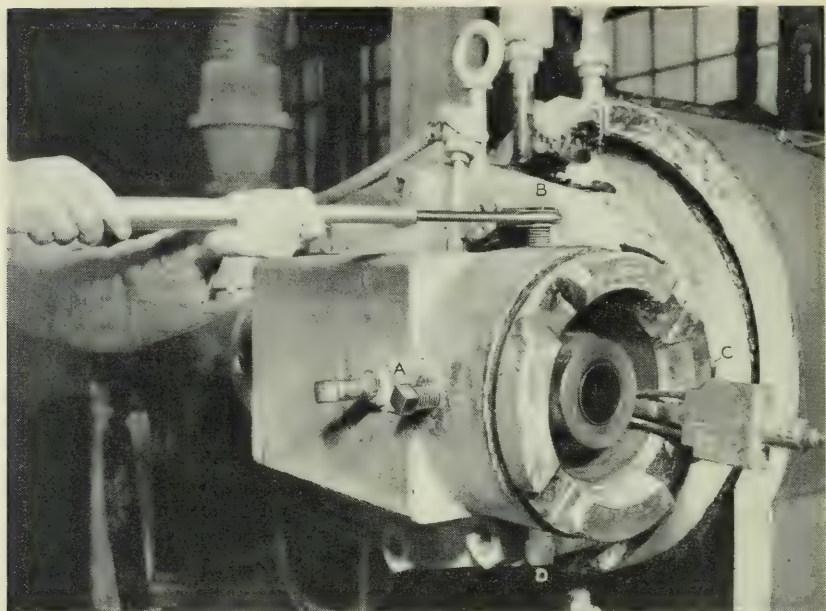


Fig. 3 — Core tube and die assembled in extruder head and die adjusting screws.

in capacitance. The capacitance response is in turn calibrated in thousandths of an inch of sheath thickness. The electronic system of the set has been described in the Bell System Technical Journal previously.* It is practical from test set measurements to control the concentricity of Alpeth cable to within 35 per cent and Stalpath to within 20 per cent. Average thicknesses within ± 0.005 inch are maintained.

Formerly, the safe practice was to use an excess of approximately 10 per cent over specified average in order to keep the thin side of eccentric sheath within the minimum spot limit. Control from test set measurements eliminated the necessity of using an excess of polyethylene because sheath of improved concentricity maintained close to the specified average thickness does not vary below the specified minimum spot thickness. The quality of the sheath is improved because it is of consistently high dimensional uniformity not previously obtainable. Also, concentric sheath has better flexing characteristics since eccentric sheath concentrates the stresses of flexing in the thin side.

* Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath, B. M. Wojciechowski, B.S.T.J., **33**, pp. 353-368, Mar., 1954.

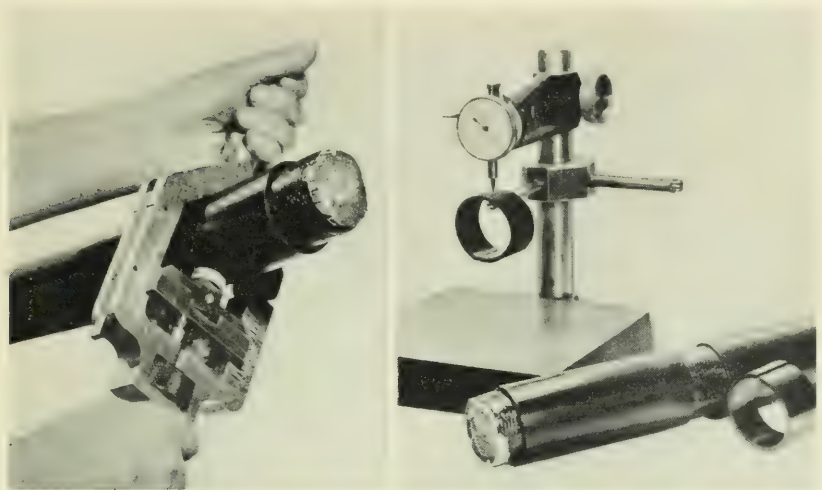


Fig. 4— (Left) Removing test strip from end of cable; (right) performing micrometer measurements on test strip.

CALIBRATION OF THE TEST SET FOR SHEATH THICKNESS MEASUREMENTS

Calibration of capacitance into thickness was difficult because the capacitance is not a simple function of polyethylene thickness. It depends also on the curvature of the sheath surface, the size and shape of the probe, the amount of flooding and the height and shape of the corrugated metal. For a given probe, it depends chiefly on the thickness, the flooding and the sheath curvature. The flooding sometimes varies from a thin film to an excess that overfills the corrugations. The surface curvature is not uniform because the soldering of the metal overlap of Stalpeth cable generally produces a flattened sector and the capstan at the soldering operation results in an elliptical shape. Changes in the surface curvature and in the amount of flooding can be compensating or cumulative in varying the capacitance.

To determine whether a correlation between jacket thickness and capacitance existed, extensive spot checks for three sizes of cable were made. Marked points on cable were measured for capacitance and then with a micrometer. A slight error can exist because the micrometer measurement is only one spot in the center of an area which is effective to capacitance. This condition is shown by Fig. 6. Also, it is difficult to determine accurately the surface curvature associated with the capacitance measurement.

The relation of thickness to capacitance conditions in the samples is

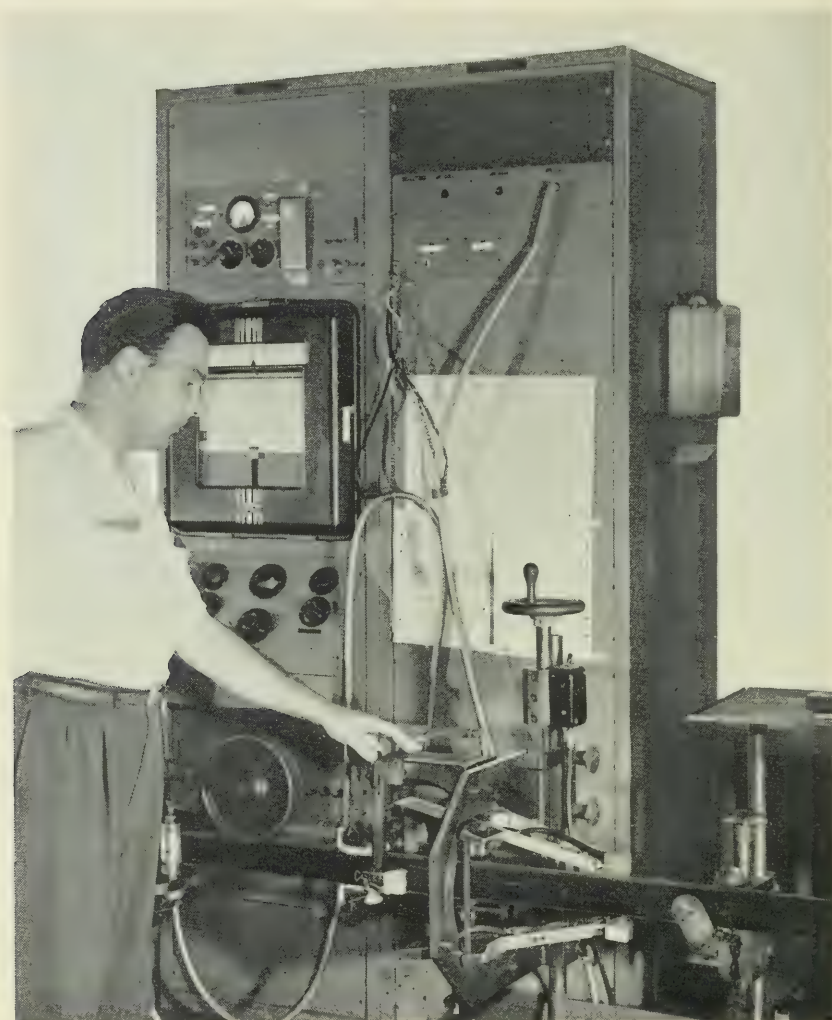


Fig. 5 — Capacitance test set and unit for tracking probes on cable surface.

shown by Fig. 7. The sheath thickness is specified as the distance between the outside surface of the sheath to the bottom of the corrugations formed into the polyethylene by the crests of the corrugated metal sheath, as indicated by dimension T . The top sketch shows the normal amount of flood. The capacitance will be different in each of the three conditions of equal thickness shown. With excess flood, center sketch, the distance between plates is increased and the capacitance is decreased.

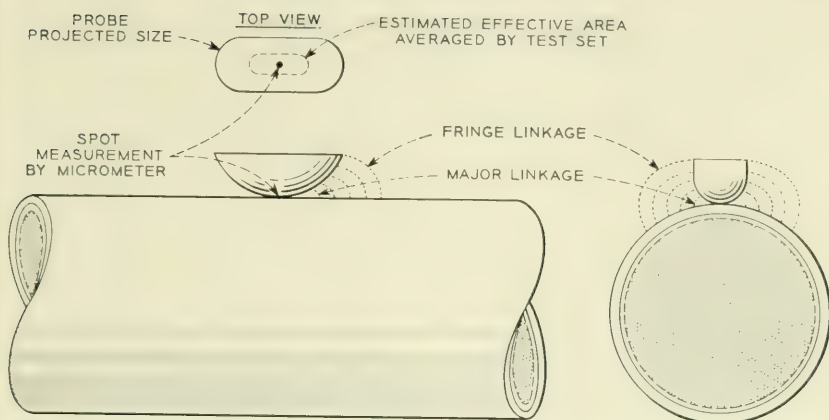


Fig. 6 — Thickness measured by direct calibration; spot by micrometer; area by capacitance.

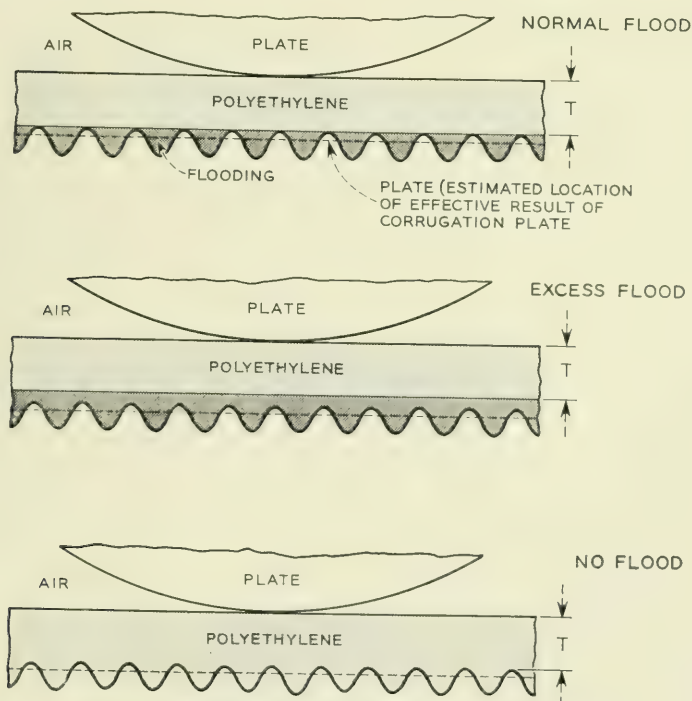


Fig. 7 — Equal thicknesses, different capacitances.

Insufficient flood, bottom sketch, alters the dielectric from polyethylene plus some flood, to all polyethylene. The capacitance is decreased.

A typical plot of points and a calibration curve are shown in Fig. 8. Each of the three cable sizes measured revealed a wide band of plot points. In each curve the points were more dense toward the left side of

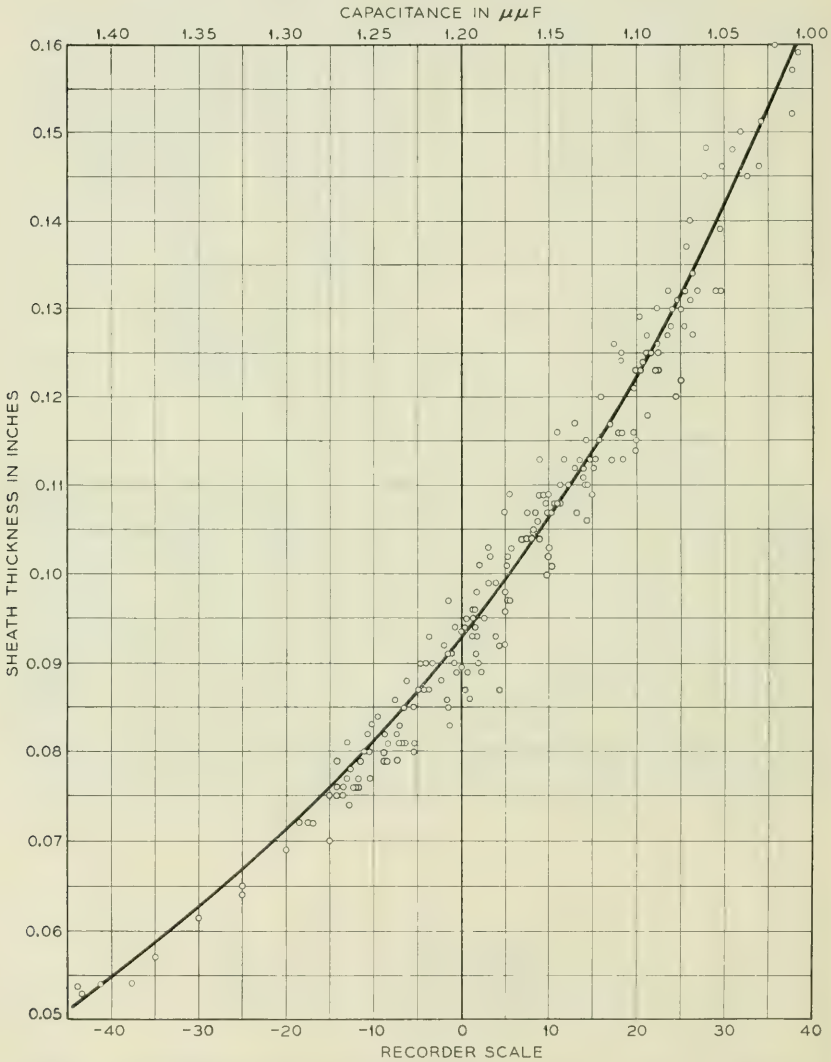


Fig. 8 — Measured points of sheath thickness versus recorder readings and developed calibration curve.

the band, becoming progressively less to the right across the band. The majority of points to the extreme right were found to be cases of excess flood. Many of the points, near the extreme right had insufficient flooding. Points close to the curve had the flood just filling the corrugation valleys. Other points consist of various other amounts of flood and/or are the result of deviation from correct surface curvature.

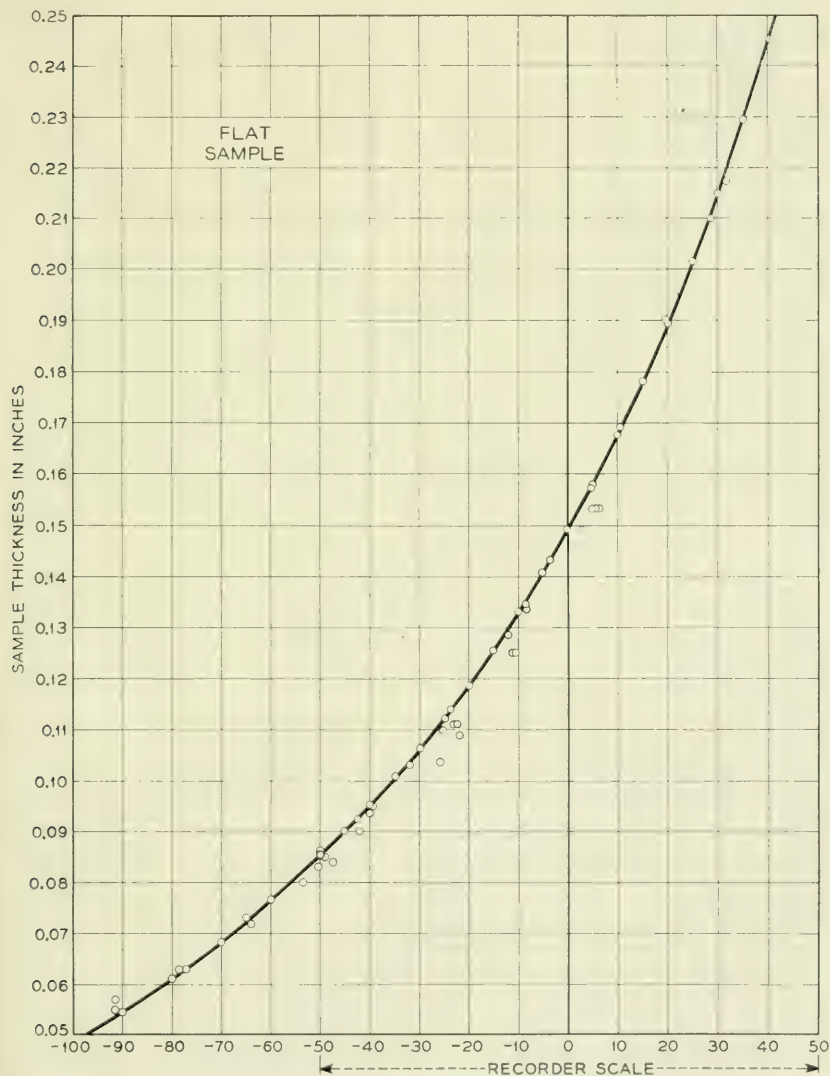


Fig. 9 — Calibration of flat sample thickness versus recorder scale.

Fig. 8 also shows that the greatest percentage of points are within a thickness range of approximately 0.010 inch. In moving downward from maximum thickness the concentration of measured thicknesses increases rapidly over approximately 0.003 inch and then becomes progressively less covering an additional 0.010 inch. The calibration curve was placed at about the location of maximum point concentration. By averaging the thickness indications along a short length of the cable, a measurement adjusted for the occasional extremes in flooding and surface variations is obtained. The accuracy for practical use is therefore within limits of ± 0.005 inch from the mean.

Investigation was also made of flat samples of Polyethylene placed upon a flat metal plate. Flat samples eliminate the variables introduced by the cable surface curvature, the corrugated metal undersheath and the flooding material. A plot of capacitance against thickness for flat samples is shown in Fig. 9. Each point represents an individual molded flat sample. The majority of points are within ± 0.003 inch of the curve.

The measurement of sufficient points to obtain curves for the many cable diameters would involve an impractical amount of work.

The calibration curves for the three cable sizes and the curve for flat samples drawn to the same capacitance versus thickness scale have similar form, but are displaced one from the other. The displacement of the calibration curves for cables of core diameters of 1.39 to 2.38 inches is shown by Fig. 10. The displacement is approximately 1 meter division for a diameter change of 0.1 inch.

Calibration curves for other cable diameters than the three measured were obtained by an approximation formula based on measuring a few points from each sheath diameter to determine the displacements and slopes and multiplying the flat sample curve values by the displacement and slope correction factors.

The curve for flat samples and the curve for 2.38 inch diameter cable plotted to the same scales is shown in Fig. 11. The two curves are sufficiently alike so that by multiplying the flat sample curve thickness values by a constant (K_1) obtained from the ratio of the cable sheath thickness to the flat sample thickness at zero recorder scale, the amount of curvature of the resultant curve and the measured sheath curve are essentially the same, and they have the same thickness and capacitance values at zero recorder reading. A multiplier (K_2) can then be added to adjust the slope of the percentage curve to make it practically coincide with the sheath thickness curve. Actually, there is a slight difference between the curvature of the flat sample curve and those of cable sheath. The amount of curvature increases as the cable diameter decreases.

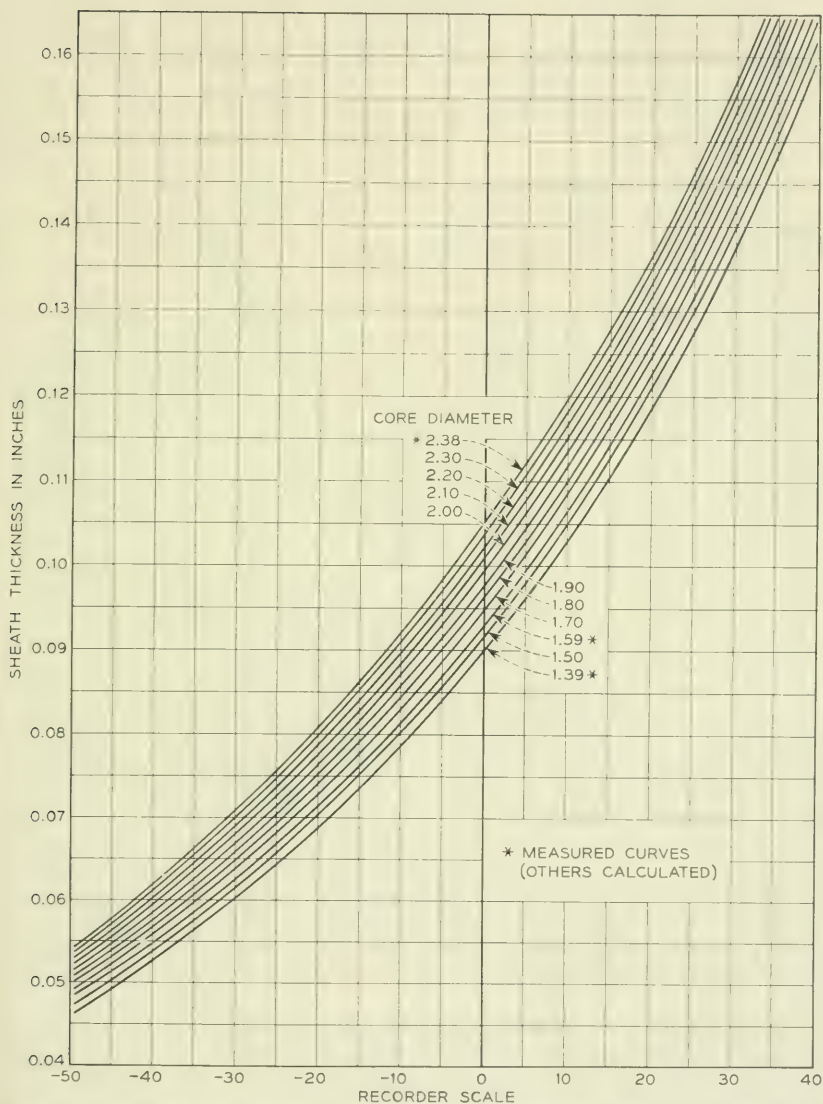


Fig. 10 — Calibration curves by core diameters, thickness versus recorder scale.

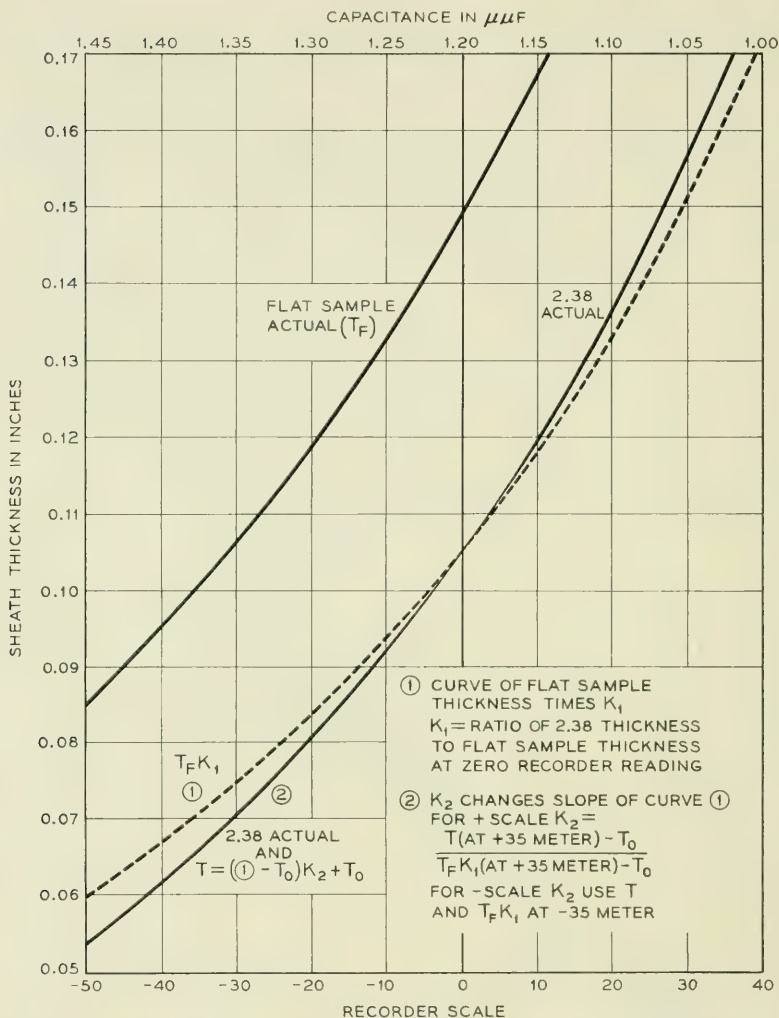


Fig. 11 — Adjustment of flat sample direct calibration to obtain calibration of 2.38-inch diameter cable sheath.

The result is the following approximation formula, from which the thickness calibration can be calculated within 0.001 inch with the error negligible over most of the working range.

$$T = T_F K_1 K_2$$

where T = Thickness in thousandth's of an inch of polyethylene cable sheath.

T_F = Thickness fo flat polyethylene sample at same recorder meter reading as for T .

K_1 = Ratio of actual cable sheath thickness to flat sample thickness at zero meter reading.

K_2 = Constant to change slope of $T_F K_1$ curve.

$$\text{For } + \text{ meter readings } K_2 = \frac{T_{(\text{at} + 35 \text{ meter})} - T_0}{T_F K_{1(\text{at} + 35 \text{ meter})} - T_0}$$

$$\text{For } - \text{ meter readings } K_2 = \frac{T_{(\text{at} - 35 \text{ meter})} - T_0}{T_F K_{1(\text{at} - 35 \text{ meter})} - T_0}$$

T_0 = Thickness in thousandth's of an inch of cable sheath at zero meter reading.

The K_1 factor accounts for the dimensional differences between the capacitor formed by a flat thickness of polyethylene on a flat plate compared to the actual capacitor construction of cable at zero meter. Both have the same capacitance of 1.20 uuF at zero meter reading. K_2 accounts for changes resulting from the curved surfaces of cable. K_1 and K_2 are different for each cable diameter.

Since zero meter is used as a reference point, the formula becomes:

$$T = (T_F K_1 - T_0) K_2 + T_0$$

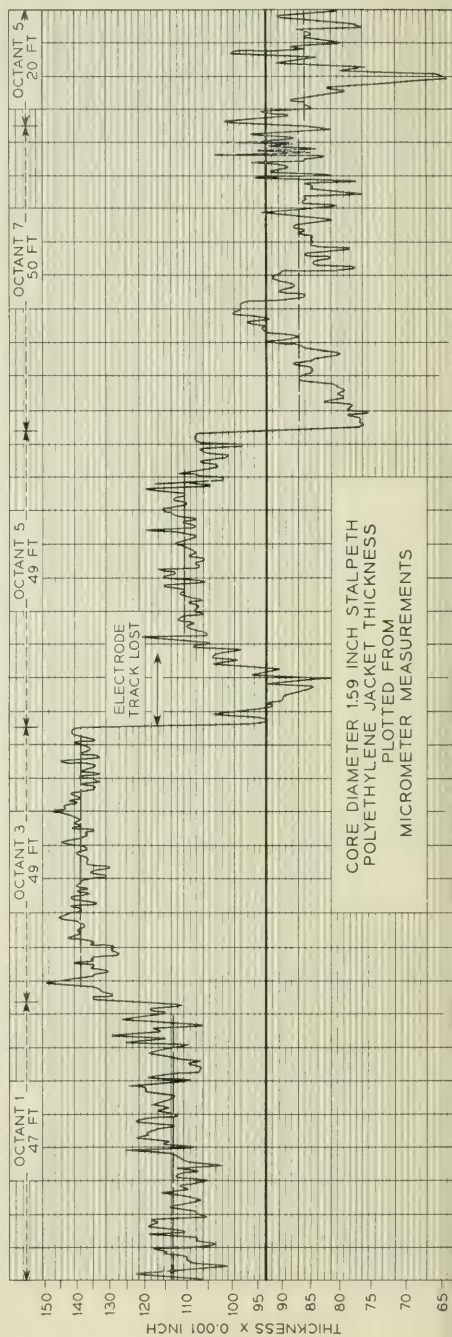
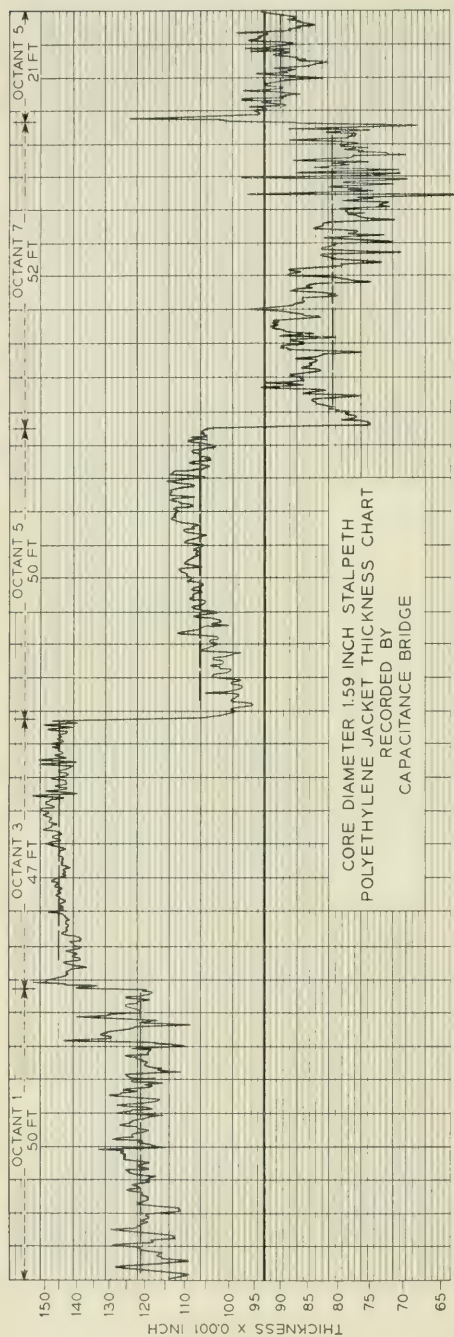
ACCURACY CHECK UNDER OPERATING CONDITIONS

A check* was made of the accuracy of calibration and of the response under operating conditions of applying the sheath to the cable. The upper graph in Fig. 12 was obtained with the test set probe tracking at a cable sheathing speed of 50 feet per minute. The probe was shifted to different octant locations on the circumference for lengths of the cable as indicated on the graphs. The track of the probe was marked on the sheath surface and the sheath then removed, cleaned of flooding compound and the micrometer measurements of the thickness taken at six-inch intervals along the length. The lower graph is a plot of the thickness obtained by micrometer. The ability of the test equipment to track and respond to the thickness variations is apparent from comparison of the two graphs.

APPLICATION OF TEST EQUIPMENT FOR EXTRUSION CONTROL

The test set is placed at some distance after the extruder to prevent the probe from marking the plastic polyethylene. The machinery of the

* Test and measurements by courtesy of J. L. O'Toole, Bell Telephone Laboratories.



sheathing line is diagrammed in Fig. 13. At the top left is the supply reel of metal jacketed cable. The cable is pulled through the flood tank where the hot rubber asphaltic compound is flowed over the corrugated metal sheath. It then progresses through the die head of the extruder where the polyethylene sheath is extruded over the flooded metal sheath. The cable with plastic polyethylene then enters the cooling trough where it is cooled and solidified. At the exit of the cooling trough is an air blower for drying the water from the sheath surface. The test set is located after the dryer. The next unit is the capstan which pulls the cable. At the final unit to the right, the sheathed cable is taken up on the shipping reel.

A typical recorder graph taken along 360 feet of cable length with the sensing probe held at one location on the sheath circumference is shown in Fig. 14. With apparently stable conditions of extrusion the spot thickness indications will vary as much as plus or minus 0.010 inch while the lengthwise average remains stable as shown in Fig. 14. These fluctuations are sheath thickness variations which result from the complex interaction of the many sheathing line variables, but they may be increased or decreased by response to uneven flooding distribution and/or variations in surface curvature. However, it is practical to visually average this graph to within ± 0.001 inch.

For die adjustment, thickness measurements are obtained visually by estimating the average of the fluctuations of the recorder's visual indicator. Measurements are taken at quadrant locations corresponding to the locations of the four die adjusting screws. Opposite thicknesses give the amount of eccentricity. Die adjustments can be made accurately because the amount of eccentricity is known and the amount of die movement is governed by the adjusting screw pitch.

Adjustment to specified average sheath thickness is made by averaging measurements at eight positions equally spaced around the sheath. Increasing the speed of the cable in relation to the speed of extrusion increases the stretch of the polyethylene and decreases the average thickness. Decreasing the cable speed increases the average thickness.

APPLICATION OF TEST EQUIPMENT FOR SHEATH INSPECTION

The thickness test provides an accurate gage for the inspection organization to measure compliance of the sheath to specified requirements. Inspection possibilities with the thickness test set are many and the problem becomes one of an economic procedure that will assure the required quality. Continuous recording of the entire cable length is practical but is unnecessary from a manufacturing viewpoint. Recorder chart speed is one half inch per minute and cable speeds are from 20 to 100

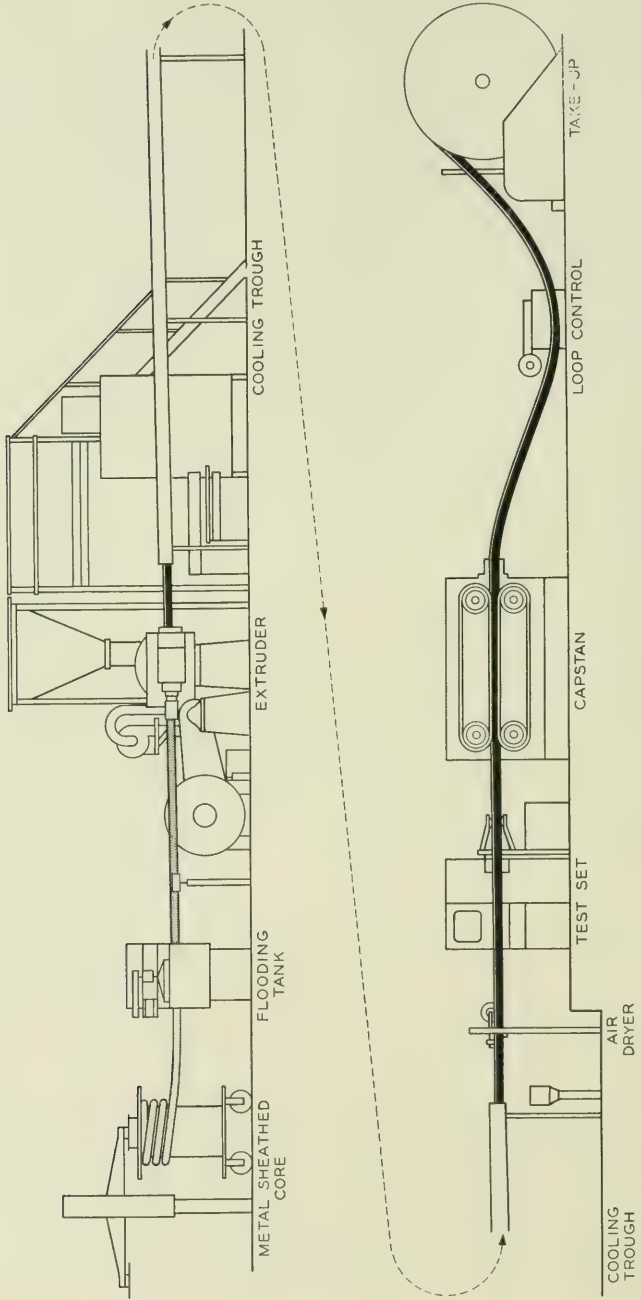


Fig. 13 — Polystyrene sheathing line.

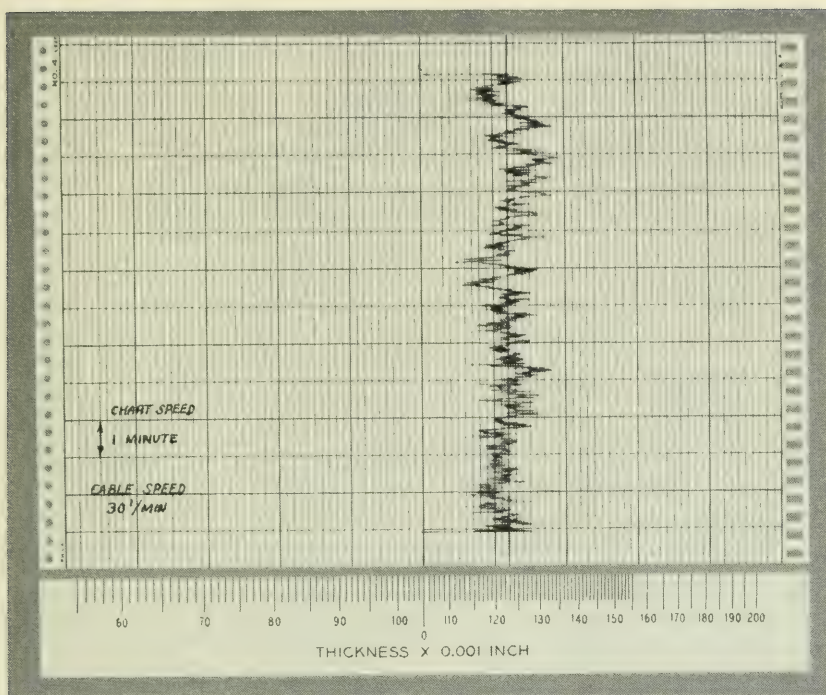


Fig. 14 — Recorder graph of single octant variation and thickness scale in thousandths of an inch.

feet per minute. It was found that the fluctuations or variation peaks of one line along the cable length could be averaged from chart lengths of $\frac{1}{4}$ inch. Also that by taking measurements consecutively by octants around the circumference a practical measure of the entire circumference is obtained and is sufficient coverage to locate the minimum wall thickness. The graphs of Fig. 15 show typical inspection recordings of two cable lengths.

Four thicknesses are specified for inspecting sheath, all of which are obtained from a graph of the consecutively recorded octants. These checks are:

1. The minimum spot thickness.
2. The average thickness lengthwise along the thinnest side. (Average of minimum octant.)
3. The average cross sectional thickness. (Average of octant averages).
4. The maximum difference between the lengthwise average of the thickest side (average of maximum octant) and the lengthwise average of the thinnest side (average of minimum octant).

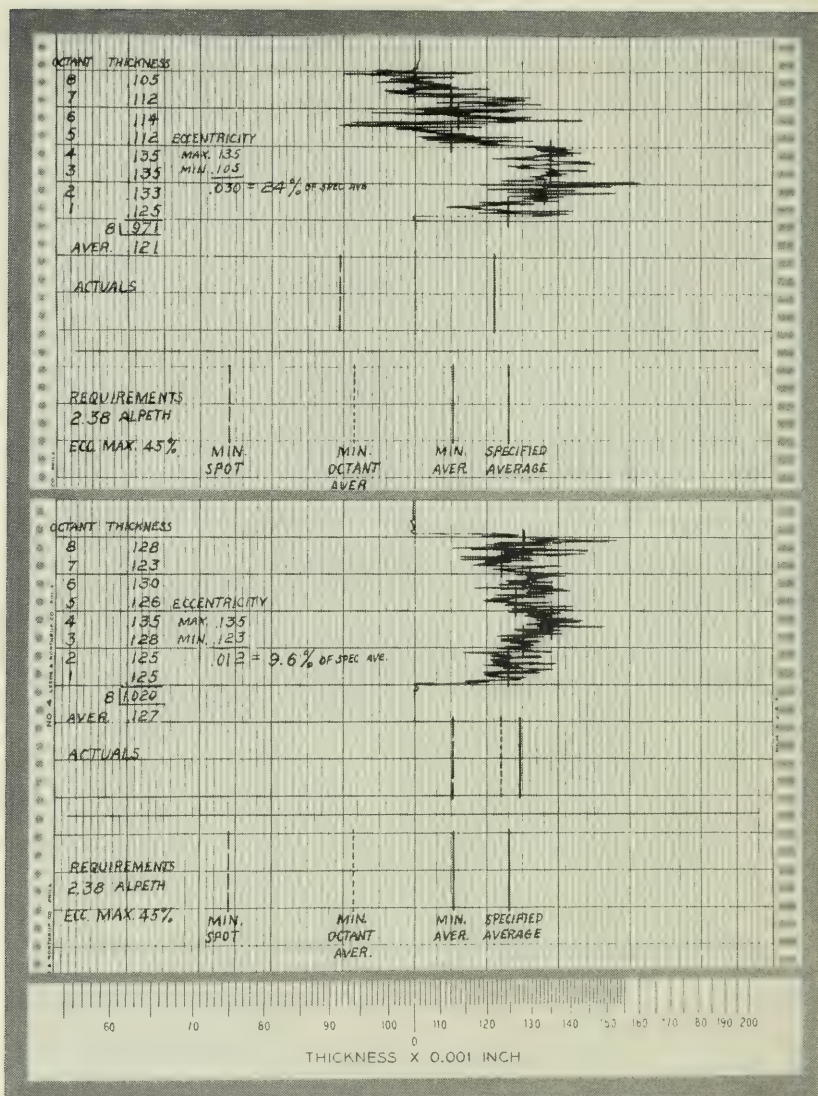


Fig. 15 — Inspection graphs of two reel lengths — octant graphs with estimated octant averages — calculation of average thickness and eccentricity; location of specified thickness and actual thickness, in thousandths of an inch.

The location of the four major thickness limits have been indicated below the test graphs.

CONCLUSIONS

This test equipment has proved to be a practical means for the control of the concentricity and the average thickness of the polyethylene sheath on Alpeth and Stalpeth cables. It is accurate, reliable and of rigid construction suitable for continuous shop use. It measures the sheath wall thickness directly in thousandths of an inch both visually and as a recorded graph and does so non-destructively as the sheath is applied.

Concentricity is maintained within 35 per cent on Alpeth and within 20 per cent on Stalpeth cable. Average thickness is controlled to within ± 0.005 inch of specified average thickness by the practice of visually averaging graphs of about twenty-five feet of cable length.

Polyethylene is conserved in two ways which reduce manufacturing costs. First, improved control permits operating at specified average thickness without varying below minimum spot limit. Previously, an excess over specified average thickness was necessary to prevent the wider range of variation from going below the specified minimum spot thickness. Second, the sheath is of consistently uniform dimensional quality not previously obtainable which made it practical to reduce the average wall thickness 11 per cent below previously specified thickness.

ACKNOWLEDGMENT

The writer wishes to express his appreciation of the co-operation of B. M. Wojciechowski of the Western Electric Company, who designed the capacitance test set and of Bell Telephone Laboratories cable engineers, in establishing the sheath requirements and for their encouragement in this project.

Topics in Guided-Wave Propagation Through Gyromagnetic Media

Part I — The Completely Filled Cylindrical Guide

By H. SUHL and L. R. WALKER

(Manuscript received January 26, 1954)

The characteristic equation for the propagation constants of waves in a filled circular guide of arbitrary radius is written in terms of magnetizing field and a carrier density, which are shown essentially to determine the dielectric and permeability tensors for a gas discharge plasma and for a ferrite. The complex structure of the spectrum of propagation constants and its dependence upon radius and the two parameters are analyzed by a semi-graphical method, supplemented by exact formulae in special regions. Thus the course of individual modes may be charted with fair accuracy.

1. INTRODUCTION

Any material medium which propagates electromagnetic disturbances possesses a local electric or magnetic structure and it is just the motion of the electric or magnetic carriers under the fields of the disturbance that determines how the propagation takes place. If a dc magnetic field be applied to the medium one may expect the local response to be altered and, consequently, to find changes in the character of the propagation. Gyromagnetic media are those for which such changes are sufficiently large to be experimentally significant. For plane waves and for optical frequencies the experimental effects and their explanation have been familiar for a great many years. The non-reciprocal rotation of the plane of polarization of light travelling parallel or antiparallel to an applied dc magnetic field, which is known as the Faraday effect, is such a phenomenon. So also is the fact that the medium becomes doubly refracting for arbitrary directions of propagation.

Interest in gyromagnetic media at longer wavelengths first arose in connection with radio propagation in the ionosphere. The ionosphere is essentially an ionic cloud and the earth supplies a magnetic field, which, for the charge densities involved, is sufficient to produce a large effect

upon propagation. Here, as in the earlier optical cases, the disturbances considered are essentially plane waves. In recent years, with the extensive development of microwave techniques, two gyromagnetic media have been investigated using guided waves. One of these is the gas discharge plasma, an ionic medium like the ionosphere, in which, however, the charge density may be varied over wide ranges in a controllable manner. The magnitude of the effects observed in such ionic media are governed by the relation of the applied frequency to the cyclotron frequency of the ions in the dc magnetic field. Goldstein and his associates¹ have studied the propagation of waves in a cylindrical waveguide within which a discharge is supported and to which a longitudinal magnetic field is applied. Among many effects which they have observed is a large Faraday rotation.

The other medium being actively investigated is the low-loss ferromagnetic medium, as exemplified by the ferrites. In this case the peculiarities of the medium have their origin in the precession of the magnetization of the ferrite about the applied field. This precession takes place with a frequency dependent upon the applied field strength and large changes in the nature of the propagation occur when the frequency of the r.f. applied field approaches this. Polder² worked out the effective properties of such a medium for plane waves and Hogan³ has made various experimental studies of the propagation in cylindrical guides containing ferrite. Here, again, Faraday rotation and other non-reciprocal effects have been observed.

In this paper a variety of topics associated with the theory of guided waves in gyromagnetic media is considered, with the main emphasis laid on the ferrites. The exposition does not attempt to be systematic. Very few problems in this field admit of a thorough analytic treatment and, frequently, the more closely allied they are to the practical uses of ferrites in microwave devices the more fragmentary is the analysis. On the other hand since the problems can always be formulated it is always possible in specific cases to resort to a purely numerical solution. The problems considered here all arise in the effort to analyze the operation of various devices and different idealizations are utilized in particular cases.

In Part I the general properties of gyromagnetic media are discussed and the connection between the phenomenological constants of the medium and the underlying molecular model is derived for the ferrite and for the plasma. The assumptions necessary to render the ferrite problem tractable are discussed at some length. Maxwell's equations are written down for a general gyromagnetic medium and some of the salient features of their solution are noted. The propagation of circularly po-

larized waves in circularly cylindrical guide filled with ferrite or plasma is then considered. The characteristic equation connecting frequency and propagation constant is first derived. For the purpose of obtaining results which can be compared with experiment, a specific molecular model is chosen for the ferrite. In this way the ferrite itself is specified by a single parameter, its saturation magnetization, and its state by another, namely the applied field. The object of the calculation, then, is to find, for a given ferrite and a given guide radius, the mode spectrum of the wave guide and the variation of propagation constant with magnetic field. This is done by a semi-graphical method supplemented by exact analytic formulae in the neighborhood of certain critical points, series expansions in certain regions and some numerical computations in others. A sketch of a similar procedure applicable to the plasma is given.

It should be pointed out that the filled cylindrical waveguide is not a topic of the highest importance from the technical standpoint. It is for this reason that no effort is made here to obtain a comprehensive body of exact numerical information about the modes. One wishes, on the other hand, to exploit the simplifying features of the problem (as contrasted with the more useful case of a cylinder of ferrite not filling the guide) so that the discussion may be exhaustive, in the sense that the complete mode spectrum is exhibited.

In Part II we deal with cases of transverse magnetization. By that term we mean the following: the microwave fields propagate in a direction normal to the dc magnetization and they do not vary along the magnetization direction. They may then be separated into two independent sets of field components, of which only one explicitly depends on the dc magnetizing field. For these two fields wave impedances are defined which can be used for matching purposes. A few simple examples are then given. One special case, that of the "non-reciprocal helix" utilizing ferrite, is of importance in traveling-wave tube work and is discussed at length.⁷ The slow-wave propagation along both a cylindrical and a "plane" helix are treated; magnetic loss is analyzed in some detail for the plane case, and general rules are given for its approximate determination in the cylindrical case.

In Part III perturbation theory and some miscellaneous topics are taken up. Suitable perturbation methods are developed for cases in which the wave guide fields are drastically modified over small volumes (as occurs if thin pencils or thin discs are inserted) and also for situations in which the local properties of the medium are but slightly disturbed over finite volumes. Among the miscellaneous topics discussed is the

propagation between infinite parallel planes filled with ferrite in a longitudinal magnetic field. The effect upon Faraday rotation of multiple reflections is considered.

2. THE PHYSICAL PROPERTIES

The propagation of electromagnetic waves in a medium is governed by Maxwell's equations which connect the space variations of $\underline{E}$ and $\underline{H}$, the electric and magnetic intensities with the time variations of $\underline{D}$ and $\underline{B}$, the electric displacement and magnetic induction. To characterize the particular medium relations may be given of the form $\underline{D} = \|\epsilon\| \underline{E}$ and $\underline{B} = \|\mu\| \underline{H}$ where $\|\epsilon\|$ and $\|\mu\|$ are the dielectric and permeability tensors. For disturbances whose amplitude is in some appropriate sense small, the elements of these tensors will be independent of rf amplitude, but will depend upon the dc state of the medium, upon the frequency of the signal and in unfavorable cases upon the wavelength of the latter. With the assumptions made in this paper the dependence upon wavelength will not arise.

The form of $\|\epsilon\|$ and $\|\mu\|$ may be known experimentally or it may be deduced from some molecular model of the medium. If the equations of motion of the parts of the medium are known under applied electric and magnetic fields, the displacement and magnetic induction resulting from this motion may be found explicitly. In isotropic media and in the absence of applied dc fields, each component of the displacement or of induction depends in the same way upon the associated component of $\underline{E}$ or $\underline{H}$. The tensors then become diagonal with equal elements. The application of a dc magnetic field, say in the z -direction, causes ions to circle about this field or magnetic dipoles to precess about it. It follows that a rf electric field in the ionic case or magnetic field in the ferrite, normal to the dc magnetic field, will produce a component of motion at right angles to itself and in time quadrature with it. From symmetry and from the equations of motion in a magnetic field the tensors may be expected to be now of the form

$$\begin{vmatrix} a & -jb & 0 \\ jb & a & 0 \\ 0 & 0 & c \end{vmatrix} \quad (1)$$

where a is an even function of magnetic field and b an odd function. c , in general, will be independent of the magnetic field.

That a and b at a given frequency and for a given sample of the medium are not independent but are related through the magnetizing dc field, H_0 , is a fact of which we need not take cognizance when solving Maxwell's

equations subject to the appropriate boundary conditions. Their solution will determine the propagation constant β of a wave as a function of a and b , no matter what their interrelation. On the other hand, in a given experiment β is generally determined as a function of one parameter only: the magnetizing field H_0 . Comparison of the family of calculated results $\beta = \beta(a, b)$, with the results $\beta = \beta(H_0)$, found experimentally will, of course, determine a and b as functions of H_0 .

If, however, we have a prior knowledge of a and b in terms of H_0 , either through postulating the correct dynamical model for the medium, or through independent experiments, we can utilize the functional form of a and b in our analysis of β , and thus arrive directly at β as a function of H_0 . The distinction between the two methods is by no means academic; early introduction of such a functional form of a and b into the waveguide problem actually simplifies the analysis. Aside from this pragmatic consideration the latter method seems to us more appropriate for another reason: it is hardly the task of analysis of technical devices to check on the physical theories that give a and b as functions of H_0 ; such checks are made by experiments specifically designed to avoid the analytic complexities attending the solutions for most of the technically important structures.

Accordingly we adopt the more direct approach of expressing a and b in terms of H_0 (and, of course, in terms of the magnetic or electric carrier density of a given sample) throughout these papers, even in those few cases in which β can be expressed analytically as a function of a and b .

2.1 Ferrites

Most ferrites used in microwave applications are fully saturated in dc magnetic fields that are small compared with the dc field with which they are biased in operation. We shall therefore always postulate a fully saturated sample. Accordingly the magnetization vector $\underline{M}$ at a point in the sample will always be of constant magnitude, although its orientation will change in the ac field.

One equation of motion for $\underline{M}$ that takes this into account is

$$\frac{d\underline{M}}{dt} = \gamma[\underline{M} \times \underline{H}_T] - \frac{\gamma\alpha}{|\underline{M}|} [\underline{M} \times [\underline{M} \times \underline{H}_T]] \quad (2)$$

where $\underline{H}_T$ is a total effective magnetic field seen by the spins that make up $\underline{M}$, t is the time and γ is the gyromagnetic ratio appropriate to electron spins, whose g -factor is close to 2. The expression on the right hand side of (2) is in the nature of a torque; the force on $\underline{M}$ is always at right angles to $\underline{M}$, thus leaving its magnitude unchanged. The first term on

the right of (2) is quite well substantiated by quantum mechanical considerations. It is a vector normal to $\underline{M}$ and to the force $\underline{H}_T$ and is responsible for the precession. The second term is also a vector normal to $\underline{M}$, but is in the plane of $\underline{M}$ and $\underline{H}$ in a sense such as to reduce the angle of the precession. It thus represents a damping. Not much is known about the precise mechanism of the damping, so that its phenomenological representation by the second term of (2) is still in doubt.

$\underline{H}_T$, the total field acting on the electron spins, is made up of terms not all of which are of electromagnetic origin. It consists of the dc field $\underline{H}_0$ within the sample, the ac field $\underline{H}$, the anisotropy field, and the field ascribed to the quantum mechanical exchange forces between spins.

$\underline{H}_0$ in the sample must be calculated from the *applied* dc field $\underline{H}_{\text{ext}}$ by a purely magnetostatic calculation, which, in the case of sufficiently simple shapes, can be carried out with the help of the appropriate demagnetizing factors. Throughout this paper it is assumed that this problem has been solved, so that $\underline{H}_0$ is given. Furthermore it is assumed that $\underline{H}_{\text{ext}}$ and $\underline{H}_0$ are uniform. Boundary effects due to non-uniformities of $\underline{H}_0$ are neglected.

The microwave field $\underline{H}$ in the sample is one of the unknowns of the problem of propagation, and will appear in the solution of Maxwell's equations subject to the appropriate boundary conditions.

The anisotropy field, a property of a single crystal of ferrite, arises from the fact that through the medium of spin-orbit interaction, the electron spins can "see" the orbital wave-functions. Since these have the symmetry properties of the crystal, it is to be expected that the anisotropy field will be a vector function of $\underline{M}$, with the symmetry properties of the crystal. The samples of ferrite used in practice contain a great many small crystals randomly oriented, so that the net effect of the anisotropy field on microwave propagation must be obtained by means of an averaging procedure. The integrations involved are laborious and have not been carried out so far. We shall therefore neglect anisotropy altogether. Since anisotropy fields are usually of the order of a few hundred gauss, this will put our results in error below frequencies of about 3,000 mc/sec. (Corresponding to a precession frequency of $\gamma H_0 = 3,000$ mc/sec., H_0 is about 1,100 gauss.)

The field between two spins ascribable to exchange forces will be zero when the two are parallel, and thus arises out of differences of spin orientation (that is, differences of $\underline{M}$) from place to place. In fact, analysis shows that this magnetic field is proportional to $\nabla^2 \underline{M}$ for cubic crystals. Thus equation (2) really involves position coordinates as well as time. Hence the ac part $\underline{m}$ of $\underline{M}$ at a point will depend not only on the ac field

$\underline{H}$ at that point, but on values of $\underline{H}$ throughout the volume of the sample. Therefore $\underline{B}$, which is $\mu_0 \underline{H} + \underline{m}$, will likewise be a functional of $\underline{H}$ over the whole sample. Fortunately it turns out that the spatial variation of $\underline{H}$ in a microwave structure is so much slower than that characteristic of the "spin waves" to which $\nabla^2 \underline{M}$ gives rise that this effect is quite negligible at microwave frequencies. Only in the most immediate vicinity of gyromagnetic resonance could such effects become significant.

Thus, we shall regard $\underline{H}_T$ simply as the sum of the dc and ac magnetic fields, $\underline{H}_0 + \underline{H}$, and correspondingly $\underline{M}$ as the sum of the dc magnetization (directed along $\underline{H}_0$ in a saturated sample when anisotropy is neglected) plus an ac part $\underline{m}$. Equation (2) must now be solved for $\underline{m}$ in terms of $\underline{H}$. It is a non-linear equation, whose solution $\underline{m}$ will depend on $\underline{H}$ non-linearly, as will $\underline{B}$. Even if $\underline{m}$ could be determined in this way, Maxwell's equations would become non-linear, and hope of their solution remote. It is therefore necessary, and in the great majority of applications also quite sufficient, to assume that the ac quantities in (2) are so small that their products can be neglected and only linear terms taken into account. The terms $\underline{m}$ and $\underline{H}$ may now be assumed to vary as $\exp j\omega t$.

Under these circumstances, (2) becomes

$$\frac{d\underline{m}}{dt} = \gamma([\underline{m} \times \underline{H}_0] + [\underline{M}_0 \times \underline{H}]) - \frac{\alpha\gamma}{|\underline{M}_0|} ([\underline{M}_0 \times [\underline{m} \times \underline{H}_0]] + [\underline{M}_0 \times [\underline{M}_0 \times \underline{H}]]),$$

and is easily solved for $\underline{m}$ in terms of $\underline{H}$, and of the dc quantities H_0 , M_0 which we shall assume to point in the z -direction. Each of the components m_x , m_y is a linear function of both H_x and H_y and when they are substituted in the components of the equation $\underline{B} = \mu_0 \underline{H} + \underline{m}$, lead to expressions of the form (1) for $\underline{B}$ in terms of $\underline{H}$:

$$\begin{aligned} B_x &= \mu H_x - j\kappa H_y, \\ B_y &= j\kappa H_x + \mu H_y, \quad \text{and} \\ B_z &= \mu_0 H_z. \end{aligned} \tag{3}$$

It is convenient to introduce two auxiliary quantities

$$\sigma = \frac{|\gamma| H_0}{\omega}; \quad p = \frac{|\gamma| M_0}{\mu_0 \omega},$$

and in terms of these one obtains the relations first derived by Polder:

$$\begin{aligned}\frac{\mu}{\mu_0} &= 1 + p \frac{\sigma(1 + \alpha^2) + j\alpha \operatorname{sgn} p}{\sigma^2(1 + \alpha^2) - 1 + 2j\alpha\sigma \operatorname{sgn} p}, \quad \text{and} \\ \frac{\kappa}{\mu_0} &= \frac{-p}{\sigma^2(1 + \alpha^2) - 1 + 2j\alpha\sigma \operatorname{sgn} p},\end{aligned}\quad (4)$$

where the function

$$\begin{aligned}\operatorname{sgn} p &= +1 & p > 0 \\ &= -1 & p < 0\end{aligned}$$

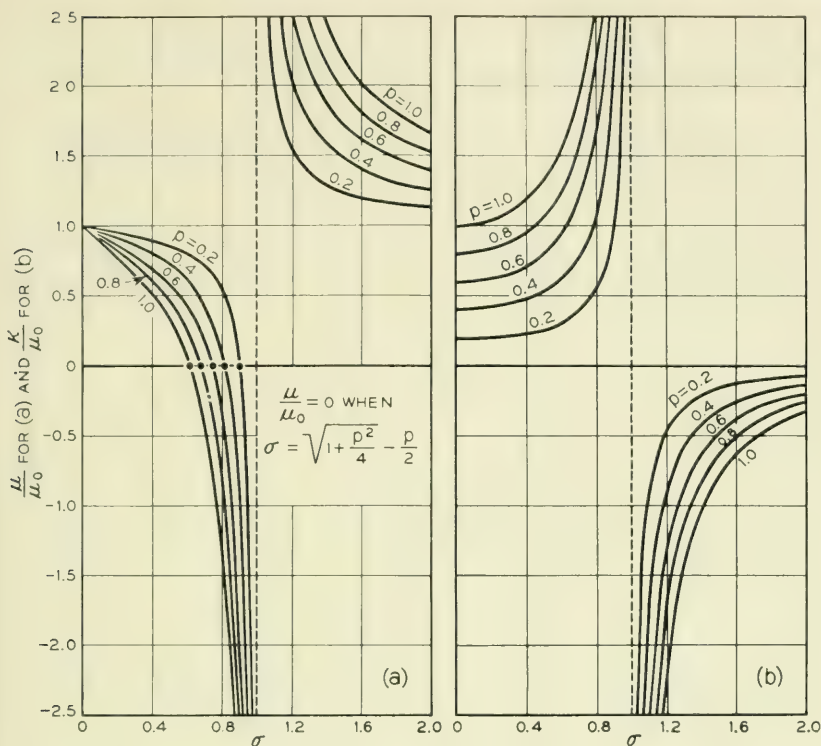
σ is the ratio of the natural precession frequency $\frac{1}{2\pi} |\gamma| H_0$ to the signal frequency. p is the ratio of a frequency $\frac{1}{2\pi} |\gamma| M_0/\mu_0$, associated with the saturation magnetization M_0 , to the signal frequency. Note that σ and p always have similar signs: if H_0 is reversed, so is the saturation magnetization. Equations (4) are true only for a fully saturated sample. Therefore they hold good only for values of σ greater than the very small value corresponding to the amount of H_0 required to saturate the sample. In practice that value of H_0 is generally so small that this restriction is trivial. In the text a number of formulae will appear which apply "near $\sigma = 0$ ". These are to be understood as applying near the very small value of σ that corresponds to saturation.

Equation (4) has an interesting implication with regard to the loss parameter α . If α were zero, we would have

$$\begin{aligned}\frac{\mu}{\mu_0} &= 1 - \frac{p\sigma}{1 - \sigma^2}, \quad \text{and} \\ \frac{\kappa}{\mu_0} &= \frac{p}{1 - \sigma^2},\end{aligned}\quad (5)$$

and these equations describe the loss-free case. If in equations (5), σ is replaced by $(\sigma + j\alpha \operatorname{sgn} p)$, the resulting expressions check (4) to order α . For small α , it follows that any propagation problem need be considered for the loss-free case (5) only.* The first order change due to loss in any formula so obtained can be deduced by differentiation of the formula

* A form of the damping term in Equation (2), no less justified experimentally than the one used above, is $-\frac{\alpha}{|\underline{M}|} \left(\underline{M} \times \frac{d\underline{M}}{dt} \right)$. When this expression is used the permeabilities are exactly functions of the variable, $\sigma + j\alpha \operatorname{sgn} p$.



Figs. 1(a) and 1(b) — The relative permeabilities μ/μ_0 and κ/μ_0 versus σ .

with respect to σ , and multiplication by $j\alpha \operatorname{sgn} p$. Of course this procedure is invalid close to resonance ($\sigma = 1$), when terms in α^2 play an important part. Equations (5), which will hereafter be called the Polder equations, are plotted in Fig. 1. The quantity

$$\rho_H = \frac{\kappa}{\mu} = \frac{p}{1 - p\sigma - \sigma^2}$$

is shown in Fig. 1(c). It occurs in the waveguide theory, and also in the theory of other microwave circuits considered later on. The ratio $\mu/\mu_0 = \mu/\mu_z$ will be denoted by ν_H . At a fixed p , μ/μ_0 decreases from unity at $\sigma = 0$, through zero at $\sigma = -p/2 + \sqrt{p^2/4 + 1}$ to $-\infty$ at $\sigma = 1 - 0$, and then from $+\infty$ at $\sigma = 1 + 0$ steadily down to unity at $\sigma = \infty$. κ/μ_0 increases from p at $\sigma = 0$ to $+\infty$ at $\sigma = 1 - 0$, and then again from $-\infty$ at $\sigma = 1 + 0$ to zero at $\sigma = \infty$.

It has already been mentioned that the anisotropy fields are of the order of a few hundred gauss. For most ferrites the saturation magnetiza-

tion is about 1,000 gauss. It will therefore be consistent with the neglect of anisotropy to assume that the applied frequency is such that p is less than unity and this will be done hereafter.

2.2 Ion clouds or plasmas

Since these are considered in much less detail in these papers, their physical properties are stated only briefly here.

Instead of a tensor relationship between $\underline{B}$ and, $\underline{H}$ we now have one between the displacement vector $\underline{D}$ and the electric field $\underline{E}$. If the magnetizing field is along the z axis, we have

$$\begin{aligned} D_x &= \epsilon E_x - j\eta E_y, \\ D_y &= j\eta E_x + \epsilon E_y, \text{ and} \\ D_z &= \epsilon_z E_z. \end{aligned} \quad (6)$$

If the medium consists of equal densities R of positive ions and electrons, and if collisions and thermal velocities are neglected, ϵ and η can be calculated for weak ac disturbances $\underline{E}e^{j\omega t}$ from the equation of motion

$$\dot{\underline{v}} = \frac{e}{m} \underline{E}e^{j\omega t} + \gamma[\underline{v} \times \underline{H}],$$

where $\underline{v}$ is the velocity vector of the electron and $\gamma = e\mu_0/m$, in the usual notation. When this equation is solved and the abbreviations

$$\omega_0 = |\gamma| H_0; \quad \sigma = \frac{\omega}{\omega_0}; \quad q = \frac{\omega_p}{\omega}; \quad \omega_p = \frac{Re^2}{m\epsilon_0}$$

are introduced, one obtains, from the fact that the total current is $j\omega\epsilon_0\underline{E} + R\underline{v}$, the heavy ions being assumed stationary,

$$\begin{aligned} \epsilon &= \epsilon_0 \left(1 + \frac{q^2}{\sigma^2 - 1} \right), \\ \eta &= \epsilon_0 \frac{q^2 \sigma}{\sigma^2 - 1}, \text{ and} \\ \epsilon_z &= \epsilon_0 (1 - q^2), \end{aligned} \quad (7)$$

where ϵ_0 is the dielectric constant of vacuum. The waveguide theory will involve the parameters

$$\begin{aligned} \nu_E &= \frac{\epsilon}{\epsilon_z} = 1 - \frac{\sigma^2 q^2}{(1 - \sigma^2)(1 - q^2)}, \text{ and} \\ \rho_E &= \frac{\eta}{\epsilon} = \frac{q^2 \sigma}{\sigma^2 + q^2 - 1}. \end{aligned} \quad (8)$$

These results apply to stationary plasmas only. If the plasma were an electron stream moving along the wave-propagation direction, for example, the dielectric constants would depend on wavelength also, and the propagation problem would be much more involved.

The variations of ϵ and η with σ are shown in Fig. 2. ϵ/ϵ_0 for a given q^2 starts at $\sigma = 0$, $\epsilon = \epsilon_0(1 - q^2)$, decreases through zero at $\sigma = \sqrt{1 - q^2}$ to $-\infty$ at $\sigma = 1 - 0$, starts again from $+\infty$ at $\sigma = 1 + 0$, and decreases to ϵ_0 at $\sigma = \infty$. $\eta = 0$ when $\sigma = 0$, decreases to $-\infty$ at $\sigma = 1 - 0$ and then decreases from $+\infty$ at $1 + 0$ to zero at $\sigma = \infty$.

We note that similar formulae apply to the electron-gas in semiconductors at temperatures sufficiently low and frequencies sufficiently high so that damping is not important. However, the formulae have to be

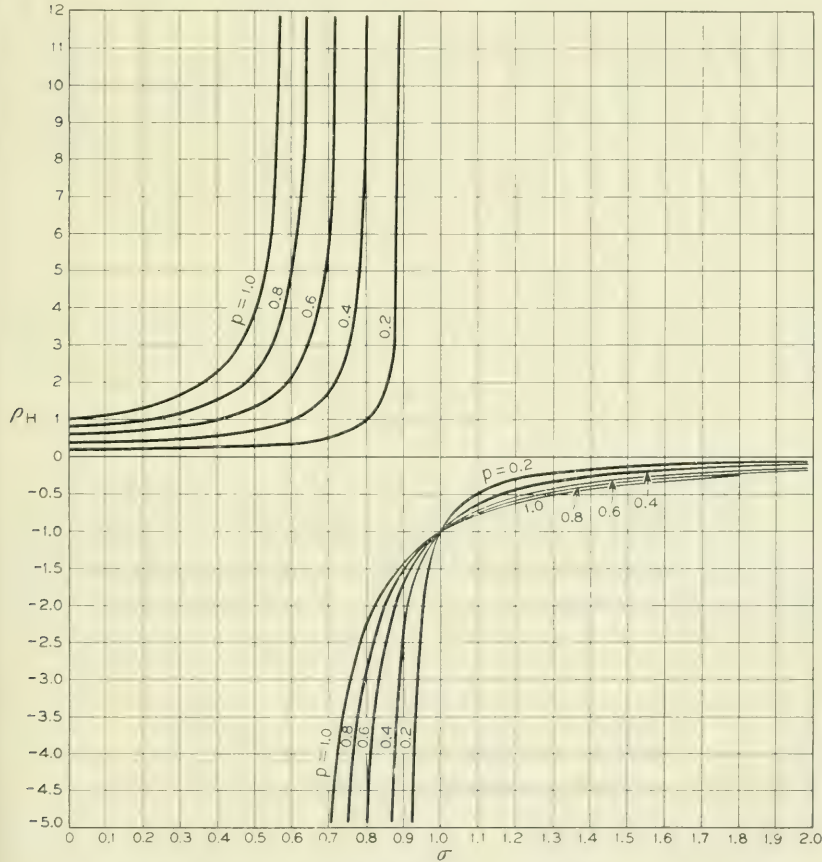


Fig. 1(c) — The ratio $\rho_H = \kappa/\mu$ versus σ .

generalized in some of those cases to take into account the existence of groups of electrons with different "effective masses" m .

3. THE SOLUTION OF MAXWELL'S EQUATIONS

Maxwell's equations will now be solved in a cylindrical waveguide filled with a hypothetical medium which contains the ferrite and the plasma as a special case. It will be supposed therefore that both its permeability and dielectric constant are tensors of the form previously considered.

3.1 Field components

The following notation will be found convenient. The projection of a vector $\underline{A}$ upon the plane normal to the z -axis will be written $\underline{A}_t$. If the components of $\underline{A}_t$ are α, β then an associated vector having components $(\beta, -\alpha)$ is devoted by $\underline{A}_t^*$. A similar notation is used for differential operators. Thus, if ∇ denotes $(\partial/\partial x, \partial/\partial y)$, ∇^* denotes $(\partial/\partial y, -\partial/\partial x)$ †. Denoting scalar products by a dot, the following identities are evident

$$\underline{A}_t^* \cdot \underline{A}_t^* = \underline{A}_t \cdot \underline{A}_t; \quad (\underline{A}_t^*)^* = -\underline{A}_t; \quad \underline{A}_t \cdot \underline{A}_t^* = 0;$$

$$\underline{A}_t \cdot \underline{B}_t^* = -\underline{A}_t^* \cdot \underline{B}_t;$$

and

$$\underline{A}_t \cdot \underline{B}_t^* = z\text{-component of } [\underline{A} \times \underline{B}].$$

Also if $\underline{k}$ is a unit vector along the positive z -axis, $\underline{k} \times \underline{A} = -\underline{A}_t^*$. Similar relations hold for differential operators. If one denotes the starring operation by the symbol P then clearly

$$P^2 = -1; \quad P^{-1} = -P; \quad \frac{1}{P+a} = \frac{1}{1+a^2} (a - P),$$

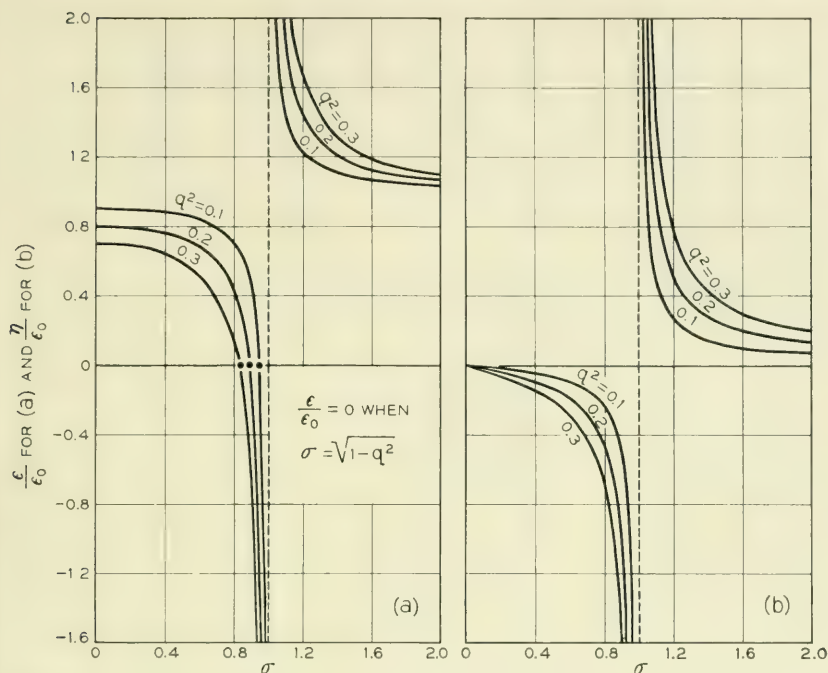
where a is a number.

Maxwell's equations may now be written, for that case in which the dependence of any component upon t and z is of the form $e^{j(\omega t - \beta z)}$, in the form:

$$\begin{aligned} \nabla^* H_z + j\beta \underline{H}_t^* &= j\omega \epsilon \underline{E}_t + \omega \eta \underline{E}_t^*, \\ \nabla \cdot \underline{H}_t^* &= j\omega \epsilon_z E_z, \\ \nabla^* E_z + j\beta \underline{E}_t^* &= -j\omega \mu \underline{H}_t - \omega \kappa \underline{H}_t^*, \text{ and} \\ \nabla \cdot \underline{E}_t^* &= -j\omega \mu_z H_z, \end{aligned} \tag{9}$$

where use is made of equations (3) and (6).

† The operator ∇^* is called "flux" by Schelkunoff. Strictly, one should write ∇_t and ∇_t^* , rather than ∇ and ∇^* , but this is needlessly cumbersome.



Figs. 2(a) and 2(b) — The relative dielectric constants ϵ/ϵ_0 and η/ϵ_0 versus σ .

It is desirable to remove scale factors as far as possible. A unit of length given by

$$\frac{1}{\beta_0} = \frac{1}{\omega \sqrt{\mu_z \epsilon_z}}$$

will be used to measure lengths. This unit is $\lambda_0/2\pi$, where λ_0 is the wavelength in an unbounded, unmagnetized medium. It will be assumed that β is in future measured in units of β_0 . Finally all magnetic fields will be multiplied by $\sqrt{\mu_z/\epsilon_z}$ to give them the dimensions of electric fields. Using the definitions of the ν 's and ρ 's given in Section 2, Maxwell's equations may be put into the form:

$$\nabla^* H_z + j\beta \underline{H}_t^* = \nu_E(j\underline{E}_t + \rho_E \underline{E}_t^*), \quad (10a)$$

$$\nabla \cdot \underline{H}_t^* = jE_z, \quad (10b)$$

$$\nabla^* E_z + j\beta \underline{E}_t^* = -\nu_H(j\underline{H}_t + \rho_H \underline{H}_t^*), \quad (10c)$$

and

$$\nabla \cdot \underline{E}_t^* = -iH_z. \quad (10d)$$

$\underline{E}_t$ and $\underline{H}_t$ may now be eliminated yielding two simultaneous second order equations for E_z and H_z . These, in turn, may be combined to produce two independent second order equations each of which is satisfied by an appropriate linear combination of E_z and H_z . These equations may be solved and E_z and H_z expressed as linear combinations of the solutions. The transverse fields are then written in terms of E_z and H_z and, finally, the boundary conditions are applied leaving a transcendental equation in β^2 .

Operating on (10a) and (10c) with $\nabla \cdot$ and taking account of (10b), (10d), one finds that

$$\begin{aligned} j\beta \nabla \cdot \underline{H}_t^* &= \nu_E(j\nabla \cdot \underline{E}_t + j\rho_E H_z) = -\beta E_z, \text{ and} \\ j\beta \nabla \cdot \underline{E}_t^* &= -\nu_H(j\nabla \cdot \underline{H}_t + j\rho_H E_z) = \beta H_z \end{aligned} \quad (11)$$

Operating on (10a) and (10c) with ∇^* , using $\nabla^* \cdot \nabla^* = \nabla^2$ and so on, one obtains, using (10b) and (10d),

$$\begin{aligned} \nabla^2 \underline{H}_z + j\beta \nabla \cdot \underline{H}_t &= \nu_E(-H_z + \rho_E \nabla \cdot \underline{E}_t), \text{ and} \\ \nabla^2 \underline{E}_z + j\beta \nabla \cdot \underline{E}_t &= -\nu_H(E_z + \rho_H \nabla \cdot \underline{H}_t). \end{aligned} \quad (12)$$

Now, elimination of $\nabla \cdot \underline{E}_t$ and $\nabla \cdot \underline{H}_t$ between (11) and (12) yields

$$\begin{aligned} \nabla^2 H_z + \nu_E \left(1 - \rho_E^2 - \frac{\beta^2}{\nu_E \nu_H} \right) H_z &= j\beta(\rho_E + \rho_H) E_z, \text{ and} \\ \nabla^2 E_z + \nu_H \left(1 - \rho_H^2 - \frac{\beta^2}{\nu_E \nu_H} \right) E_z &= -j\beta(\rho_E + \rho_H) H_z, \end{aligned} \quad (13)$$

equations which demonstrate that pure *TE* or *TM* fields no longer exist, as the result of the presence of ρ 's. H_z or E_z might now be eliminated between these equations giving a single equation in ∇^2 and $(\nabla^2)^2$, but it is more convenient to find those linear combinations of E_z and H_z which satisfy a first order equation in ∇^2 . Writing such a linear combination as

$$\psi = E_z + j\Lambda H_z, \quad (14)$$

and adding $j\Lambda$ times the first of equations (13) to the second, it is found that this is an equation in ψ alone of the form

$$\nabla^2 \psi + \chi^2 \psi = 0, \quad (15)$$

provided that Λ is a root of the quadratic

$$\Lambda^2 - \frac{\nu_E \left(1 - \rho_E^2 - \frac{\beta^2}{\nu_E \nu_H} \right) - \nu_H \left(1 - \rho_H^2 - \frac{\beta^2}{\nu_E \nu_H} \right)}{\beta(\rho_E + \rho_H)} \Lambda - 1 = 0. \quad (16)$$

The value of χ^2 is then given by

$$\chi_{1,2}^2 = \nu_E \left(1 - \rho_E^2 - \frac{\beta^2}{\nu_E \nu_H} \right) - \beta(\rho_E + \rho_H) \Lambda_{2,1}, \quad (17a)$$

or

$$\chi_{1,2}^2 = \nu_H \left(1 - \rho_H^2 - \frac{\beta^2}{\nu_E \nu_H} \right) + \beta(\rho_E + \rho_H) \Lambda_{1,2}, \quad (17b)$$

where Λ_1 and Λ_2 are the roots of (16) and χ_1^2, χ_2^2 are the corresponding χ^2 . The labelling of the roots is not important, but consistency must be maintained. From (14) E_z and H_z must satisfy

$$E_z + j\Lambda_1 H_z = \psi_1,$$

and

$$E_z + j\Lambda_2 H_z = \psi_2$$

so that

$$E_z = \frac{\Lambda_2 \psi_1 - \Lambda_1 \psi_2}{\Lambda_2 - \Lambda_1}, \quad (18a)$$

and

$$H_z = j \frac{\psi_1 - \psi_2}{\Lambda_2 - \Lambda_1}. \quad (18b)$$

Solutions of (15) may now be sought in cylindrical coordinates. To satisfy the boundary conditions in circular guide it will be necessary to assume the solutions to vary as $e^{jn\theta}$, where θ is the polar angle and n is any integer, positive, negative or zero. Equation (15) then becomes

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \psi_{1,2}(r)}{\partial r} \right) + \left(\chi_{1,2}^2 - \frac{n^2}{r^2} \right) \psi_{1,2}(r) = 0,$$

if r is the radius. Solutions which are regular within the guide will have the form of constant multiples of $J_n(\chi_{1,2}r)$, where J_n is the n^{th} order Bessel function. The solutions of (15) are, then,

$$\psi_{1,2} = A_{1,2} J_n(\chi_{1,2}r) e^{jn\theta}, \quad (19)$$

where the A 's are constants. E_z and H_z can be found now from (18), but further equations must be found to express $\underline{E}_t$ and $\underline{H}_t$. Using P to denote the starring operation, (10a) and (10c) may be re-written as

$$(j\nu_F P - \rho_E \nu_E) \underline{E}_t + j\beta \underline{H}_t = -\nabla \underline{H}_z,$$

and

$$-j\beta \underline{E}_t + (j\nu_H P - \rho_H \nu_H) \underline{H}_t = \nabla E_z,$$

which yield

$$\begin{aligned} \{[\nu_E \nu_H (1 + \rho_E \rho_H) - \beta^2] - j\nu_H \nu_E (\rho_H + \rho_E) P\} \underline{E}_t \\ = -j\beta \nabla E_z - \nu_H (jP - \rho_H) \nabla H_z, \end{aligned}$$

and

$$\begin{aligned} \{[\nu_E \nu_H (1 + \rho_E \rho_H) - \beta^2] - j\nu_H \nu_E (\rho_H + \rho_E) P\} \underline{H}_t \\ = \nu_E (jP - \rho_E) \nabla E_z - j\beta \nabla H_z \end{aligned}$$

The term in parentheses may be removed by using the rule for inverting such expressions in P which was given earlier. This process gives

$$\begin{aligned} \Omega \underline{E}_t = \left[\left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) + j(\rho_H + \rho_E) P \right] \\ [-j\beta \nabla E_z - \nu_H (jP - \rho_H) \nabla H_z], \end{aligned} \quad (20a)$$

and

$$\begin{aligned} \Omega \underline{H}_t = \left[\left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) + j(\rho_H + \rho_E) P \right] \\ [\nu_E (jP - \rho_E) \nabla E_z - j\beta \nabla H_z], \end{aligned} \quad (20b)$$

where

$$\begin{aligned} \Omega = \nu_E \nu_H \left[\left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right)^2 - (\rho_E + \rho_H)^2 \right], \\ = \nu_E \nu_H \left[\frac{\beta^2}{\nu_E \nu_H} - (1 + \rho_E)(1 + \rho_H) \right] \left[\frac{\beta^2}{\nu_E \nu_H} - (1 - \rho_E)(1 - \rho_H) \right]. \end{aligned}$$

It may be noted that for plane waves in the unbounded medium along the z axis, which have $E_z = H_z = 0$, Ω must vanish and that the propagation constants for such plane waves are evidently given by

$$\beta^2 = \nu_E \nu_H (1 \pm \rho_E)(1 \pm \rho_H). \quad (21)$$

The values of E_z and H_z given by (18) may now be substituted in (20) and the operator P removed. This gives, finally,

$$\begin{aligned} (\Lambda_1 - \Lambda_2) \Omega \underline{E}_t = j \left[\left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) (\beta \Lambda_2 - \rho_H \nu_H) + \nu_H (\rho_H + \rho_E) \right] \nabla \psi_1 \\ - \left[(\beta \Lambda_2 - \rho_H \nu_H) (\rho_E + \rho_H) + \nu_H \left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) \right] \nabla^* \psi_1 \end{aligned} \quad (22a)$$

minus the same expression with suffixes 1 and 2 interchanged.

$$\begin{aligned}
(\Lambda_1 - \Lambda_2)\Omega H_t = & \left[\left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) (\Lambda_2 \nu_E \rho_E - \beta) - \nu_E \Lambda_2 (\rho_E + \rho_H) \right] \nabla \psi_1 \\
& - j \left[\nu_E \Lambda_2 \left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) - (\rho_E + \rho_H) (\Lambda_2 \nu_E \rho_E - \beta) \right] \nabla^* \psi_1
\end{aligned} \quad (22b)$$

minus the same expression with suffixes 1 and 2 interchanged.

Equations (22a) and (22b) may be written in a variety of equivalent forms by making use of the relations between Λ_1 and Λ_2 . The manipulations which have been used in deriving (22a) and (22b) assume the use of rectangular coordinates, but the results are valid in polar coordinates if $\underline{E}_t$ means (E_r, E_θ) and ∇ means $\left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \theta} \right)$. That this is the case may be seen from the consideration that the rotation, $-\theta$, which carries the vector (E_x, E_y) into the vector (E_r, E_θ) also transforms $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)$ into $\left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \theta} \right)$.*

3.2 The characteristic equation

The boundary conditions of the problem are that $E_z = 0$ and $E_\theta = 0$ at $r = r_0$, the radius of the guide. E_z is given by [see (18)].

$$(\Lambda_2 - \Lambda_1)E_z = [\Lambda_2 A_1 J_n(\chi_1 r) - \Lambda_1 A_2 J_n(\chi_2 r)] e^{jn\theta}, \quad (23)$$

and vanishes at $r = r_0$ if

$$A_1 = \frac{J_n(\chi_2 r_0)}{\Lambda_2}; \quad A_2 = \frac{J_n(\chi_1 r_0)}{\Lambda_1}.$$

Hence the relations hold:

$$\psi_{1,2} = \frac{1}{\Lambda_{2,1}} J_n(\chi_{2,1} r_0) J_n(\chi_{1,2} r) e^{jn\theta}.$$

From (22a) it follows that

$$\begin{aligned}
(\Lambda_1 - \Lambda_2)\Omega E_\theta = & \frac{J_n(\chi_2 r_0) e^{jn\theta}}{\Lambda_2} \left[-\frac{n}{r} \left\{ \left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) (\beta \Lambda_2 - \rho_H \nu_H) \right. \right. \\
& + \nu_H (\rho_H + \rho_E) \left. \right\} J_n(\chi_1 r) + \left\{ (\beta \Lambda_2 - \rho_H \nu_H) (\rho_E + \rho_H) \right. \\
& \left. \left. + \nu_H \left(1 + \rho_E \rho_H - \frac{\beta^2}{\nu_E \nu_H} \right) \right\} \chi_1 J_n'(\chi_1 r) \right]
\end{aligned}$$

* In Appendix III the field components in polar coordinates are written out fully for the ferrite and plasma cases with some changes in notation which are introduced in Sections 4.11 and 4.2.

minus the same expression with the suffixes interchanged. Hence

$$\begin{aligned}
 & (\Lambda_1 - \Lambda_2)\Omega E_\theta(r_0) \\
 &= \frac{J_n(\chi_1 r_0)J_n(\chi_2 r_0)e^{jn\theta}}{r_0} \left[(\Lambda_1 - \Lambda_2)n\nu_H \left(\frac{\rho_H\beta^2}{\nu_E\nu_H} + \rho_E(1 - \rho_H^2) \right) \right. \\
 & \quad + \left\{ \beta(\rho_E + \rho_H) - \Lambda_1\nu_H \left(1 - \rho_H^2 - \frac{\beta^2}{\nu_E\nu_H} \right) \right\} \frac{\chi_1 r_0 J_n'(\chi_1 r_0)}{J_n(\chi_1 r_0)} \\
 & \quad \left. - \left\{ \beta(\rho_E + \rho_H) - \Lambda_2\nu_H \left(1 - \rho_E^2 - \frac{\beta^2}{\nu_E\nu_H} \right) \right\} \frac{\chi_2 r_0 J_n'(\chi_2 r_0)}{J_n(\chi_2 r_0)} \right], \quad (24)
 \end{aligned}$$

where use has been made of the relation $\Lambda_1\Lambda_2 = -1$. Therefore the characteristic equation for β^2 is obtained by equating the term in square brackets to zero. Because of the quadratic relation satisfied by Λ and the relation between Λ and χ , it is possible to write the characteristic equation in a great variety of ways. It will be convenient to introduce a function

$$F_n(x) = F_{-n}(x) = \frac{xJ_n'(x)}{J_n(x)}. \quad (25)$$

Using the F -function and replacing the Λ 's by χ 's the characteristic equation may be written:*

$$\begin{aligned}
 n\nu_H(\chi_2^2 - \chi_1^2) \left[\frac{\rho_H\beta^2}{\nu_E\nu_H} + \rho_E(1 - \rho_H^2) \right] \frac{1}{\beta(\rho_E + \rho_H)} &= \frac{\chi_2^2}{\Lambda_2} F_n(\chi_1 r_0) \\
 &- \frac{\chi_1^2}{\Lambda_1} F_n(\chi_2 r_0). \quad (26)
 \end{aligned}$$

The asymmetry of this equation between ρ_H , ν_H and ρ_E , ν_E arises from the fact that the boundary conditions involve electric field components alone.

It may be noted that if the basic solution had been taken to vary as $\cos n\theta$ or $\sin n\theta$, the expression for E_θ would have been a linear combination of $\sin n\theta$ and $\cos n\theta$ that could not have vanished at the walls for all θ .

In passing we remark that for a guide of arbitrary cross-section, the

* The characteristic equations given in Reference 4 were specializations to the ferrite and plasma cases of the form in square brackets. They have also been derived by Kales⁵ and Gamo⁶. These authors have given expressions for some, though not all, of the varieties of cut-off point derived in this paper and classified them as TE or TM according to the field configuration at cut-off. By contrast, they are classified here by their association with quasi-TE or quasi-TM limit modes which reduce to the usual TE and TM modes in the unmagnetized medium.

boundary value problem may be put into the form, of which (26) is a special case,

$$\nabla^2 f_1 + \chi_1^2 f_1 = 0,$$

and

$$\nabla^2 f_2 + \chi_2^2 f_2 = 0,$$

$$j \left(\frac{\chi_2^2}{\Lambda_2} \frac{\partial f_1}{\partial N} - \frac{\chi_1^2}{\Lambda_1} \frac{\partial f_2}{\partial N} \right) = \frac{\nu_H (\chi_2^2 - \chi_1^2)}{\beta (\rho_E + \rho_H)} \left[\frac{\rho_H \beta^2}{\nu_E \nu_H} + \rho_E (1 - \rho_H^2) \right] \frac{\partial f_1}{\partial S},$$

where $\partial/\partial N$ and $\partial/\partial S$ are normal and tangential derivatives at the guide surface, where, in addition, $f_1 = f_2$.

4. DISCUSSION OF THE PROPAGATION CONSTANTS

At this point we specialize the characteristic equation (26) to one or other of the two media.

4.1. The ferrite ($\rho_E = 0$, $\nu_E = 1$)

4.1.1. After some rearrangement the characteristic equation becomes

$$\frac{1}{\chi_1^2} \left[\frac{F_n(\chi_1 r_0)}{\lambda_1} - n \right] = \frac{1}{\chi_2^2} \left[\frac{F_n(\chi_2 r_0)}{\lambda_2} - n \right], \quad (27)$$

where $\lambda_{2,1} = \beta \Lambda_{1,2}$ and the λ satisfy

$$\lambda_{1,2}^2 - \frac{(1 - \nu_H) \left(1 - \frac{\beta^2}{\nu_H} \right) + \nu_H \rho_H^2}{\rho_H} \lambda_{1,2} - \beta^2 = 0. \quad (28)$$

The χ 's are given by

$$\chi_{1,2}^2 = \left(1 - \frac{\beta^2}{\nu_H} \right) - \rho_H \lambda_{1,2}. \quad (29)$$

From Polder's equations for ρ_H and ν_H , (28) may be written

$$\lambda_{1,2}^2 - [p + \sigma(1 - \beta^2)] \lambda_{1,2} - \beta^2 = 0, \quad (30)$$

or

$$\lambda_1 \lambda_2 = -\beta^2, \quad (31a)$$

$$\begin{aligned} \lambda_1 + \lambda_2 &= p + \sigma(1 - \beta^2), \\ &= p + \sigma + \sigma \lambda_1 \lambda_2. \end{aligned} \quad (31b)$$

If β^2 be eliminated between equations (28) and (29), $\chi_{1,2}^2$ may be ex-

pressed solely in terms of $\lambda_{1,2}$, ρ_H and ν_H in the form

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \frac{\nu_H}{\rho_H} \lambda_{1,2}}.$$

Again using Polder's formulae, this becomes

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \sigma \lambda_{1,2}}. \quad (32)$$

With these expressions for the χ , the characteristic equation takes the form

$$G(\lambda_1, \sigma, r_0) = G(\lambda_2, \sigma, r_0), \quad (33)$$

where

$$G(\lambda, \sigma, r_0) = \frac{1 - \sigma \lambda}{1 - \lambda^2} \left[\frac{1}{\lambda} F_n \left(r_0 \sqrt{\frac{1 - \lambda^2}{1 - \sigma \lambda}} \right) - n \right]. \quad (34)$$

Equations (31b) and (33) may now be considered for a fixed σ and p as determining associated pairs of values for λ_1 and λ_2 . Such a pair in turn determines $\beta^2 = -\lambda_1 \lambda_2$. Since β^2 must be positive for propagation λ_1, λ_2 must have opposite signs. The convention will be adopted that λ_1 is positive and λ_2 is negative. Equation (31b) will hereafter be called the Polder relation and Equation (33) the G -equation.

An important fact of which frequent use will be made is that the transformation

$$\lambda_1 \rightarrow -\lambda_2, \quad \lambda_2 \rightarrow -\lambda_1, \quad \sigma \rightarrow -\sigma, \quad p \rightarrow -p$$

leaves the Polder relation and β^2 unchanged and converts n to $-n$ in the G -equation. It follows that it is necessary to consider positive n only, provided we allow the pair σ, p to take on negative as well as positive values. This corresponds to the physical fact that a right-circular wave in a backward-directed magnetizing field behaves like a left-circular wave in a forward field.

The discussion in this paper is confined to the first azimuthal mode number $n = \pm 1$. Accordingly the symbol F will replace F_1 in what follows.

Before commencing the graphical analysis of the G function it is advantageous to consider briefly the function $F(x) = xJ_1'(x)/J_1(x)$, which we require for real and for purely imaginary x . By logarithmic differentia-

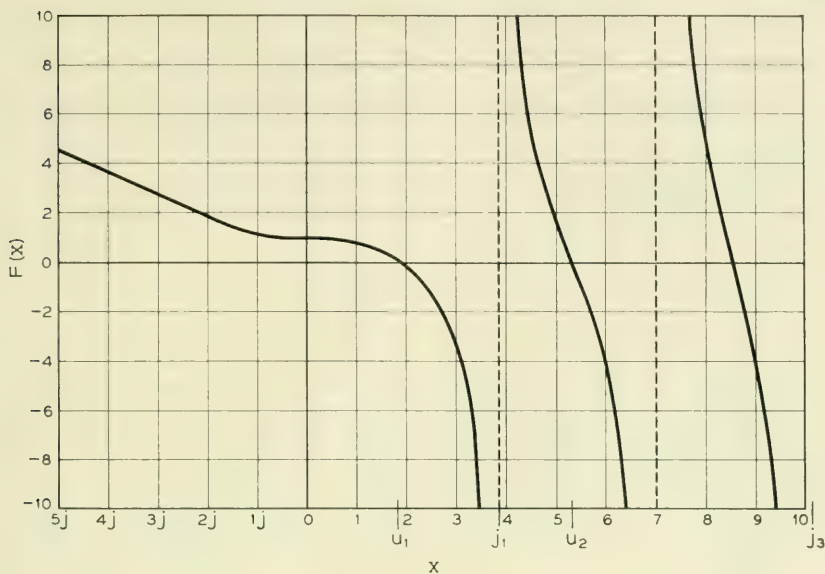


Fig. 3 — The function $F(x) = \frac{xJ_1'(x)}{J_1(x)}$.

tion of the infinite product for $J_1(x)$, $F(x)$ is found to be given by

$$F(x) = 1 - 2 \sum_{n=1}^{\infty} \frac{x^2}{j_n^2 - x^2},$$

where the j_n 's are the zeros of $J_1(x)$.

Thus, $F(x)$ is real if x^2 is, which is always the case here. For positive x , $F(x)$ is an always decreasing function of x , which has an infinite number of first order zeros and poles. The zeros are those of $J_1'(x)$ and will be denoted by u_n . The poles are the zeros of $J_1(x)$. It may be recalled from the properties of Bessel functions that for large n these zeros and poles are essentially equally spaced with a separation $\pi/2$. When x is a pure imaginary, equal to jy , $F(x)$ becomes $yI_1'(y)/I_1(y)$. This is a steadily increasing function of y , always positive, and behaving like $y - 1/2$ for large y . The function F is shown in Fig. 3. Further formulae pertaining to F are given in Appendix I. The inverse function $F^{-1}(x)$, which is also of some importance, is a multivalued function of x , whose behavior is readily understood from the figure for $F(x)$. We are now ready to proceed with the graphical analysis of the G -equation.

In a rectangular coordinate system with λ as abscissa and σ as ordinate, a contour map is sketched of the function

$$G = \frac{1 - \sigma\lambda}{1 - \lambda^2} \left[\frac{1}{\lambda} F \left(r_0 \sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}} \right) - 1 \right]$$

for all values of λ, σ from $-\infty$ to $+\infty$, r_0 being kept fixed. This can be done as accurately as desired by first drawing the contours $\chi^2 = \frac{1 - \lambda^2}{1 - \sigma\lambda} = \text{constant}$ (Fig. 4), along each of which G simply behaves like $A/\lambda + B$ and is easily evaluated with the help of a table of F . However,



Fig. 4 — Curves of constant $\chi^2 = \frac{1 - \lambda^2}{1 - \sigma\lambda}$ in the $\sigma - \lambda$ plane.

many features of the G -contours are already determined by the position of the contours $G = \infty$, and $G = 0$, across which G changes sign (from $\pm \infty$ to $\mp \infty$ or from ± 0 to ∓ 0). Because of their special role in the subsequent analysis it is desirable to introduce a scheme for their enumeration. The infinity and zero curves in the right-hand half-plane will be denoted by I and O , respectively, those in the left-hand half-plane by I' and O' . All but two of the I -curves arise from the poles j_n of F . Their equations are

$$\frac{1 - \lambda^2}{1 - \sigma\lambda} = j_n^2 / r_0^2 \quad n = 1, 2, \dots$$

Each of these curves has two branches, one in the half-plane $\lambda > 0$, one in $\lambda < 0$ and these are called I_n , I_n' respectively. All I_n curves pass through $\lambda = 1$, $\sigma = 1$, all I_n' curves pass through $\lambda = -1$, $\sigma = -1$. The lines $\lambda = 0$, $\lambda = -1$ are also infinity curves to be denoted by I_A , I_B' respectively (As $\lambda \rightarrow \pm 1$, G tends to a finite value).

Zero curves of G are given by

$$F\left(r_0 \sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}}\right) = \lambda,$$

or in a more readily computable form by

$$\sigma = \frac{1}{\lambda} - \frac{r_0^2(1 - \lambda^2)}{\lambda[F^{-1}(\lambda)]^2}. \quad (35)$$

The branches of $F^{-1}(\lambda)$ may be labelled according to the scheme: "0" for $-\infty < [F^{-1}(\lambda)]^2 < j_1^2$; "1" for $j_1^2 < [F^{-1}(\lambda)]^2 < j_2^2$ and so on. The n th branch of $F^{-1}(\lambda)$ gives rise to an O_n curve for $\lambda > 0$ and to an O_n' curve for negative λ . All O_n' curves pass through $\lambda = -1$, $\sigma = -1$; all save one of the O_n curves pass through $\lambda = 1$, $\sigma = 1$. The exceptional one, seen to be O_0 , is associated with the "0" branch of $F^{-1}(\lambda)$ on which $F^{-1}(1) = 0$. For fixed σ , G tends to zero as $\lambda \rightarrow \infty$, hence the vertical lines $\lambda = \pm \infty$ are also zero curves, to be denoted by O_∞ and O_∞' respectively.

In a sense the two branches of $\sigma\lambda = 1$ are also zero curves, to be called O_c and O_c' . O_c and O_c' are zero curves only when viewed from "one side." In the right half-plane, for $\lambda < 1$ as $\sigma\lambda \rightarrow 1 - 0$ and for $\lambda > 1$ as $\sigma\lambda \rightarrow 1 + 0$, the argument of F tends to infinity and remains real. Therefore G passes through all values an indefinite number of times and $\sigma\lambda = 1$ is a limit line of all contours, $G = \text{constant}$. For $\lambda < 1$ as $\sigma\lambda \rightarrow 1 + 0$ and for $\lambda > 1$ as $\sigma\lambda \rightarrow 1 - 0$, the argument of F is

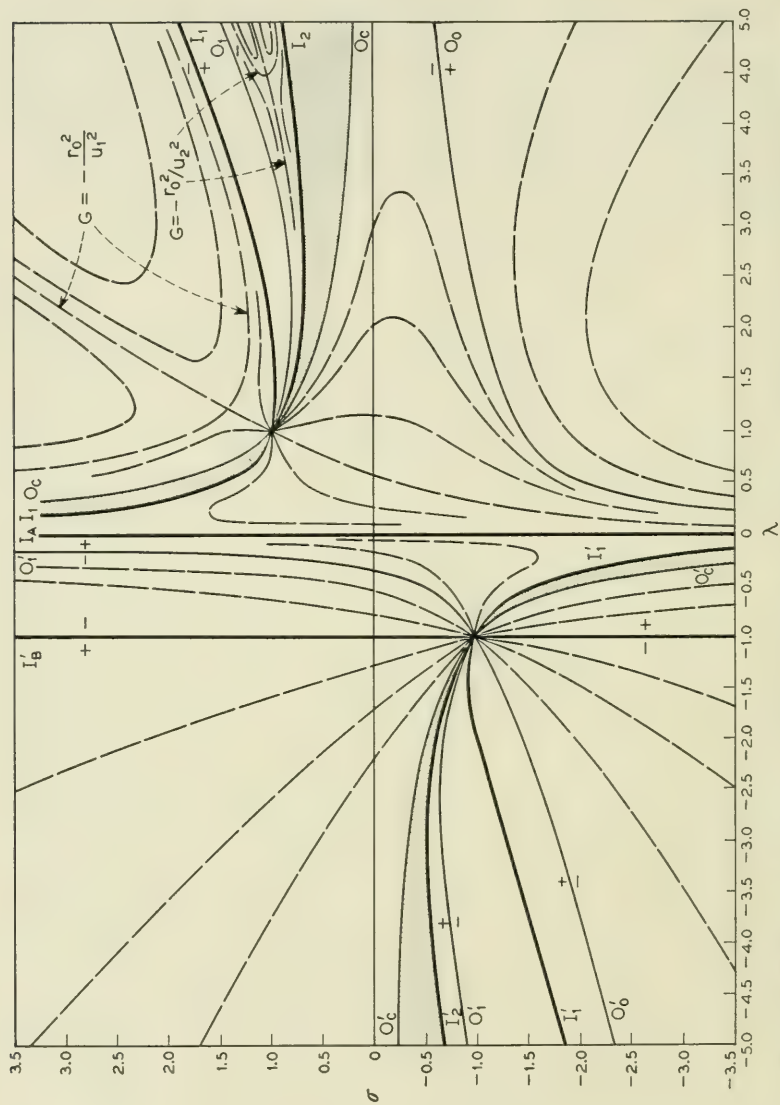


Fig. 5 — Contours of constant $G(\lambda, \sigma)$ for $\tau_0 = 2.2$. Scales distorted to show saddle points. G assumes any given value infinitely often in the shaded regions.

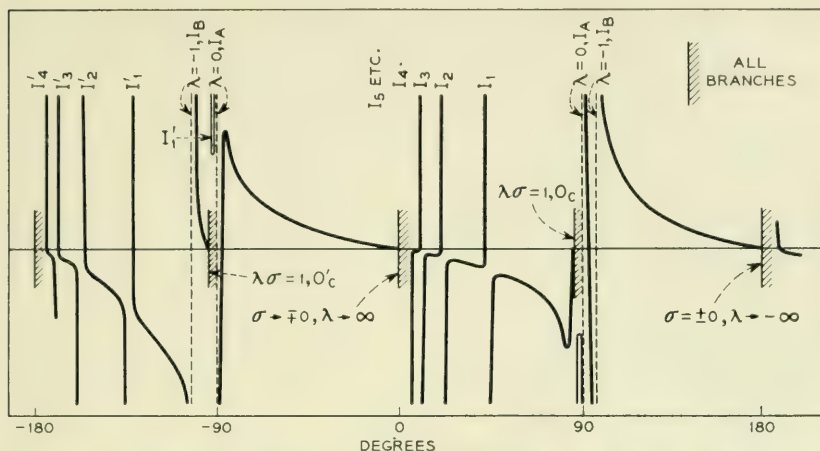


Fig. 6 — Qualitative behavior of $G(\lambda, \sigma)$ at large distances from the origin as a function of $\arctan \sigma/\lambda$. r_0 is about 2.

imaginary and F tends to infinity. However $F / \frac{1 - \lambda^2}{1 - \sigma\lambda}$ tends to zero so that G tends to zero.

To complete the picture of the G -function given by the form and position of the 0 and I curves it is necessary to see how it behaves at large distances from the origin. This is indicated in Fig. 5 and also by Fig. 6. The latter shows the value of G at large distances as a function of direction. In general, along the line $\sigma = c\lambda + d$ (c finite), G will tend to $-c$ for all d . For $c = r_0^2/j_n^2$ (which is the slope of the asymptotes to the I_n curves), G again tends to a constant. Now, however, the constant depends upon d and assumes all values from $-\infty$ to $+\infty$ as a function of d . In the first quadrant the sign of variation of the limiting value of G with direction c is opposite to that of its variation with d near $c = r_0^2/j_n^2$. Consequently local maxima and minima arise as a function of direction between successive I_n -curves. This suggests the existence of saddle points, which may be verified directly. In the third quadrant, the dependence of G upon c and d does not give such maxima and minima, and indeed no saddle points are found there. Finally it is necessary to consider the behavior of G as σ tends to infinity, while λ remains finite, corresponding to $(1/c) \rightarrow 0$. If λ remains fixed, then for $\lambda > -1$, $G \rightarrow \mp \infty$ as $\sigma \rightarrow \pm \infty$; and for $\lambda < -1$, $G \rightarrow \pm \infty$ as $\sigma \rightarrow \pm \infty$. As $\lambda \rightarrow 0$, the curves of constant G are asymptotic to $\lambda\sigma = \left(1 - \frac{r_0^2}{u_n^2}\right) - B\lambda$, where B goes from $-\infty$ to $+\infty$ with G . Interleaved with these families of curves are the curves

$G = \pm \infty$, which are $\lambda\sigma = \left(1 - \frac{r_0^2}{j_n^2}\right) + 0(\lambda^2)$. More detailed information on these matters will be found in Appendix II.

From the G -diagram it would be possible to determine pairs of λ -values with opposite signs, which, for a definite σ -value satisfy the characteristic equation, but, for a given p such pairs would not necessarily satisfy the Polder relation (31b). It is necessary to have a procedure which takes account of the latter systematically. Such a method may be based upon the fact that if, for σ and p positive, the Polder relation is solved for λ_1 in terms of λ_2 it can be thought of as a rather simple mapping of the whole λ_2 -quadrant upon a part of the λ_1 -quadrant ($\lambda_1 > 0$). Similarly for σ and p negative there is an analogous mapping of the λ_1 -quadrant onto the λ_2 -quadrant.

Considering first the case $\sigma, p > 0$, the Polder relation may be written in the forms

$$\lambda_1 = \frac{\sigma + p - \lambda_2}{1 - \sigma\lambda_2} = \frac{1}{\sigma} + \frac{\sigma + p - 1/\sigma}{1 - \sigma\lambda_2} = T(\lambda_2). \quad (36)$$

From (36) it may be seen that the curves $\lambda_2 = \text{const.}$ transform into a bundle of hyperbolae passing through the intersection of $\sigma = 1/\lambda_1$ and $\sigma = \lambda_1 - p$; that is, through λ_{10}, σ_0 , where

$$\sigma_0 = -p/2 + \sqrt{\frac{p^2}{4} + 1}, \quad \lambda_{10} = p/2 + \sqrt{\frac{p^2}{4} + 1}.$$

These hyperbolae have the vertical asymptotes $\lambda_1 = -1/\lambda_2$, and intersect $\sigma = 0$ at $\lambda_1 = p - \lambda_2$. For a fixed positive σ less than σ_0 , λ_1 decreases from $1/\sigma$ to $\sigma + p$ as λ_2 increases from $-\infty$ to 0, but when σ is greater than σ_0 , λ_1 increases from $1/\sigma$ to $\sigma + p$ under the same circumstances. Thus the whole λ_2 -quadrant is transformed upon that part of the λ_1 -quadrant which lies between the hyperbola $\lambda_1 = 1/\sigma$ and the straight line $\lambda_1 = \sigma + p$. It follows that points in the λ_1 -quadrant which are, for a given p , excluded from this region, cannot be the site of acceptable solutions of the G -equation.

Since as has already been stated, the Polder relation is unchanged by the substitution $\lambda_1 \rightarrow -\lambda_2$, $\lambda_2 \rightarrow -\lambda_1$, $\sigma \rightarrow -\sigma$, and p to $-p$, it follows that for σ and p negative a similar mapping of the λ_1 -quadrant upon part of the λ_2 -quadrant takes place. The transforms of the lines $\lambda_1 = \text{const.}$ and so forth may easily be found by using these substitutions in the formulae already given.

Reference to Fig. 1(a) and (b) will show that $\pm\sigma_0$ are the values of σ at which μ reverses sign. Therefore we may expect σ_0 to play a special role

in the propagation theory, as also does $\sigma = 1$. The following scheme exists: for $0 < \sigma < \sigma_0$, κ and μ are both positive; for $\sigma_0 < \sigma < 1$, $\kappa < 0$ and $\mu < 0$, for $\sigma > 1$, κ is negative and μ positive. If σ is changed to $-\sigma$, μ goes into μ , and κ into $-\kappa$.

The procedure which will now be used to discuss the solution of the characteristic equation, observing the Polder relations, begins by writing the equation, for σ , p positive, in the form

$$G(\lambda_1, \sigma, r_0) = G(T(\lambda_1), \sigma, r_0)$$

We are already in possession of a contour map of the left hand side of this equation in the quadrant $\sigma > 0$, $\lambda > 0$, and of the function $G(\lambda_2, \sigma, r_0)$ in the quadrant $\lambda < 0$, $\sigma > 0$. The latter surface has now to be transformed into one in the λ_1 -quadrant by the relation

$$\lambda_2 = T(\lambda_1) = (\sigma + p - \lambda_1)/(1 - \sigma\lambda_1)$$

(or equally well, $\lambda_1 = T(\lambda_2)$). This may be effected by considering the transformation of curves $G(\lambda_2, \sigma, r_0) = \text{constant}$, onto the λ_1 -quadrant. For the I' curves whose analytical expression in terms of σ and λ_2 is very simple, the corresponding explicit expression of the transformed curve in λ_1 and σ is simple. Contours other than I' are most easily transformed by replotting $G(\lambda_2, \sigma, r_0) = \text{const.}$ in the hyperbola-mesh formed by the lines $T(\lambda_2)$. However, information about particular points and about asymptotic behavior of these transformed curves is available in analytic form and is stated in Appendix II. The two surfaces so obtained will intersect in various curves, along whose projections on the $\lambda - \sigma$ plane both Polder relation and G -equation are satisfied. For each such projection λ_1 is a function of σ , λ_2 is then known in terms of σ and p , and finally $\beta^2 = -\lambda_1\lambda_2$ is known. In most cases the general course of these curves can be found without resort to much numerical analysis. Each of the curves is associated with a definite mode and it follows that the classification of the modes can be carried out fairly easily. The approximate location of the solution curves relies upon the fact that if the position of the infinity curves of both surfaces is known, continuity considerations will frequently assure the existence of an intersection within certain regions. Moreover, the neighborhood of certain special points on these solution curves can be investigated analytically. These are points at which one or both of the G -functions may be approximated by a simpler expression; included among these is the point at infinity.

It is clear that for σ and p negative the whole procedure outlined above may be carried out in a similar way, with the $\sigma > 0$, $\lambda > 0$ quadrant now being transformed on to the $\sigma < 0$, $\lambda < 0$ quadrant.

It is possible to translate such solution curves into $\beta^2 - \sigma$ curves in a direct graphical manner if a mesh of constant β^2 lines is drawn in the first quadrant. From (30) these are given by

$$\lambda_1^2 - [p + \sigma(1 - \beta^2)]\lambda_1 - \beta^2 = 0,$$

or

$$\frac{1 - \lambda_1^2}{1 - \left[\sigma + \frac{p}{1 - \beta^2} \right] \lambda_1} = 1 - \beta^2.$$

The contour, $\beta = b$, is just the contour $\chi^2 = 1 - b^2$ displaced along the σ -axis by an amount, $-p/(1 - b^2)$. The contours of constant β all pass through the point $\sigma = \sigma_0$, $\lambda = \sigma_0 + p$ and are shown in Fig. 7. Their

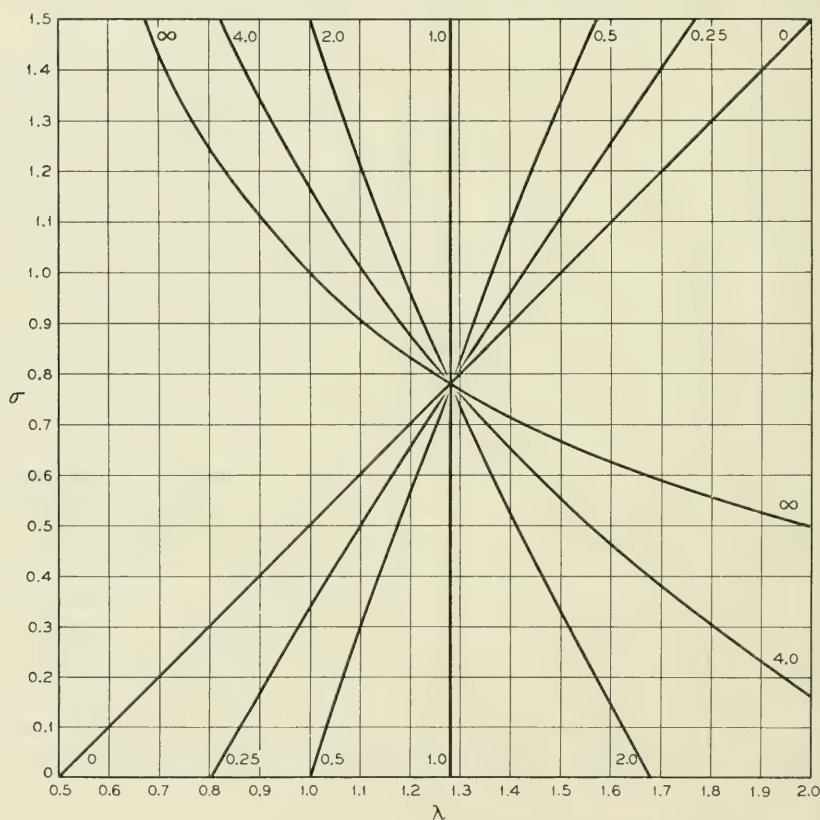


Fig. 7 — Contours of constant $\beta^2 = \lambda \frac{\lambda - p - \sigma}{1 - \sigma\lambda}$ for $p = 0.5$.

$$[\beta^2(\lambda, \sigma, p) = \beta^2(-\lambda, -\sigma, -p)]$$

course in the third quadrant is immediately found by reflection in the origin.

When $p = 0$ the magnetization of the ferrite vanishes and it is clear that we should then obtain just the modes of a guide filled with isotropic material ($\mu = \mu_z$; $\kappa = 0$). Superficially it might appear that, since the equations (31b) and (33) depend upon σ , even for $p = 0$, this result might not be attained. We now show that β^2 is indeed independent of σ for $p = 0$. It may first be noted that in this case if $\sigma \neq 1$, the Polder relation (31b) transforms

$$\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1} \text{ into } \frac{1 - \lambda_2^2}{1 - \sigma\lambda_2} \text{ and } \lambda_1 \text{ into } \frac{\sigma - \lambda_2}{1 - \sigma\lambda_2}.$$

The G -equation reads

$$\begin{aligned} \frac{1 - \lambda_2\sigma}{1 - \lambda_2^2} \left[\frac{1}{\lambda_2} F \left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \right) - 1 \right] \\ = \frac{1 - \lambda_2\sigma}{1 - \lambda_2^2} \left[\frac{1 - \sigma\lambda_2}{\sigma - \lambda_2} F \left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \right) - 1 \right]. \end{aligned}$$

Since $\lambda_1 \neq \lambda_2$ we must have

$$F \left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \right) = 0 \quad \text{or} \quad \infty,$$

or

$$\frac{1 - \lambda_{1,2}^2}{1 - \sigma\lambda_{1,2}} = \frac{u_n^2}{r_0^2} \quad \text{or} \quad \frac{j_n^2}{r_0^2}.$$

Thus $\lambda_{1,2}$ are roots of

$$\lambda^2 - \sigma \frac{u_n^2}{r_0^2} \lambda + \left(\frac{u_n^2}{r_0^2} - 1 \right) = 0,$$

or else of

$$\lambda^2 - \sigma \frac{j_n^2}{r_0^2} \lambda + \left(\frac{j_n^2}{r_0^2} - 1 \right) = 0,$$

In the first case

$$-\lambda_1\lambda_2 = \beta^2 = 1 - \frac{u_n^2}{r_0^2},$$

and in the second

$$-\lambda_1\lambda_2 = \beta^2 = 1 - \frac{j_n^2}{r_0^2}.$$

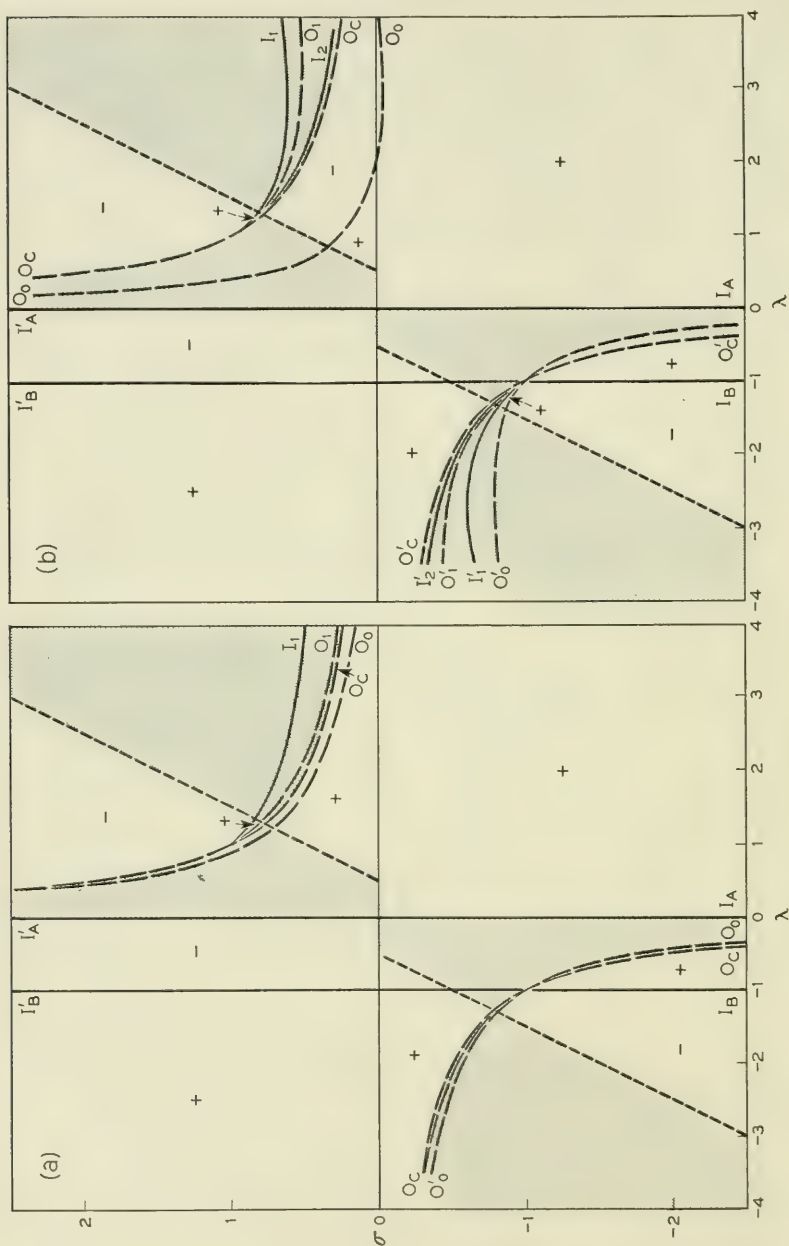


Fig. 8.— The division of the $\lambda - \sigma$ plane allowed by the Polder relation, Equation (31b), into regions of positive and negative G by the first few 0 and I curves. Excluded regions are shaded. $|p|$ is about 0.5. Fig. 8(a), $0 < r_0 < 1$; Fig. 8(b), $1 < r_0 < n_1$; Fig. 8(c), $r_0 = n_1$; Fig. 8(d), $n_1 < r_0 < j_1$; Fig. 8(e), $j_1 < r_0 < n_2$.

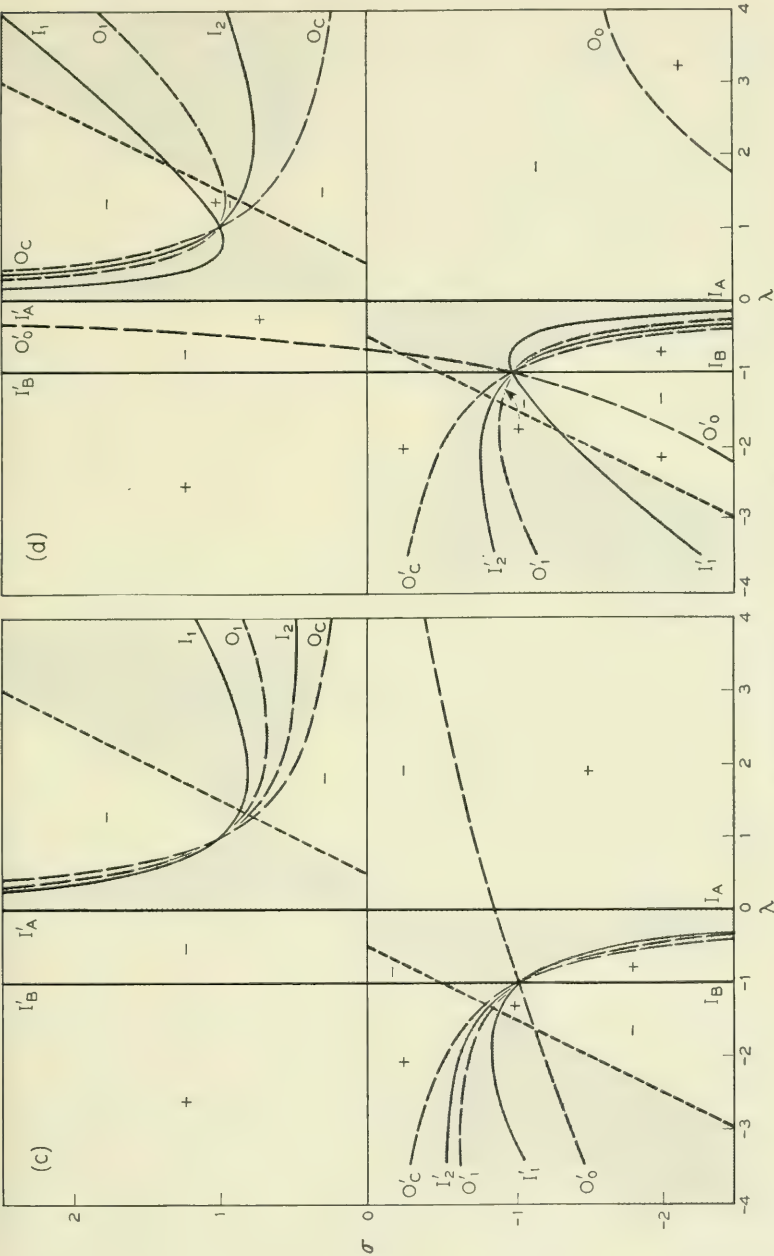


Fig. 8(c) and (d) — See Fig. 8.

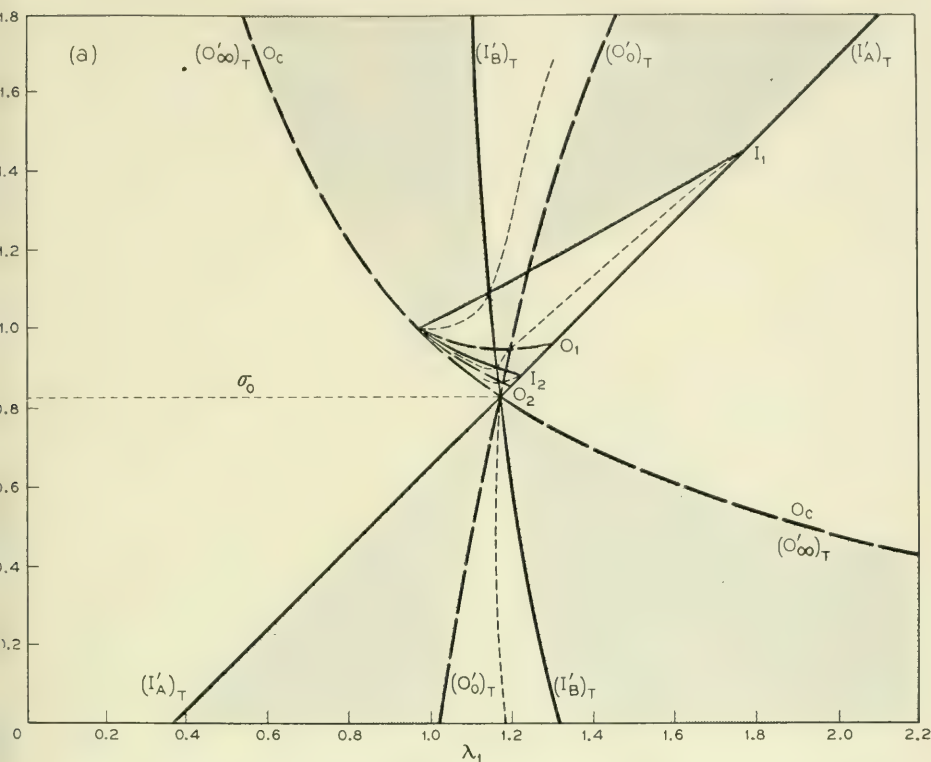


Fig. 9 — Geometrical exploration of the solution curves. The permitted areas of the $\lambda - \sigma$ plane are divided by the O , I , $(O)_T$, $(I)_T$ curves into regions in which $G(\lambda, \sigma)$ and $G(T(\lambda), \sigma)$ have like or unlike signs. Shaded regions are those of unlike signs. Solution curves (shown schematically by dotted lines) must lie in regions of like signs. Only the first few O and I curves are shown. Fig. 9(a) and (b), $r_0 \sim 3.0$; Fig. 9(c) and (d), $r_0 \sim 5.0$; Fig. 9(e) and (f), $u_2 < r_0 < j_2$; Fig. 9(g), $r_0 < u_1$. $|p| < 1$, throughout. The horizontal dashed line marks $|\sigma| = |\sigma_0|$.

Polder relation is indicated, for $p = \frac{1}{2}$, by shading. For other p -values the straight portion of the boundary of the excluded region is simply translated along the λ -axis.

In Fig. 9(a) the allowed region of the first quadrant is shown again, together with the transforms $(I'_A)_T$, $(O'_0)_T$, and $(I'_B)_T$ of the only critical curves I'_A , O'_0 and I'_B occurring in the second quadrant for the present radius. Regions in which $G(\lambda_1, \sigma)$ and $G(T(\lambda_1), \sigma)$ have opposite sign are shaded; the common signs in the remaining parts of the quadrant are as indicated. (In this diagram p is taken to be $\frac{3}{8}$).

From the disposition of the surfaces $G(\lambda_1, \sigma)$ in the region between O_1

and I_1 and of the surface $G(T(\lambda_1), \sigma)$ between $(0_\infty')_T$ and $(I_B')_T$ it is evident that along any contour such as $G(\lambda_1, \sigma) = K$, $G(T(\lambda_1), \sigma)$ will take on all positive values from 0 to ∞ and, in particular, K . Since this is true for any K , it follows that the region between 0_1 , I_1 , $(I_B')_T$ contains a solution curve. Two points on this curve are immediately obvious: the intersections of $(I_B')_T$, I_1 and the point $(1, 1)$ on $(0_\infty')_T$, 0_c . The first is the intersection of the curves

$$\lambda = 1 + \frac{p}{1 + \sigma}, \quad \frac{j_1^2}{r_0^2} = \frac{1 - \lambda^2}{1 - \sigma\lambda}.$$

At the point $(1, 1)$, λ_2 is $-\infty$, λ_1 is unity and β^2 is therefore infinite. Armed with this knowledge we now investigate analytically the behavior of β^2 near $\sigma = 0$ directly from the original G -equation and Polder relations. Writing $\lambda_1 = 1 + c\epsilon$ and $\sigma = 1 + \epsilon$, $(1 - \lambda^2)/(1 - \sigma\lambda)$ is to zero

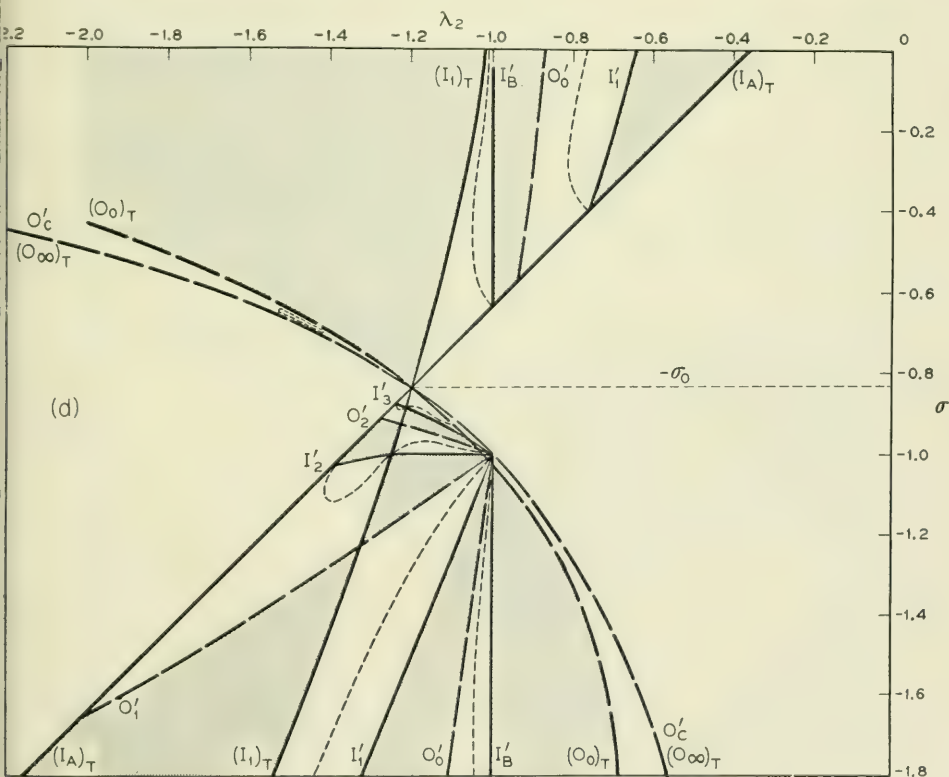


Fig. 9(d) — See Fig. 9.

function through all possible values of the other in the region bounded by I_1 , $(I_B')_T$, $(O_1')_T$ we deduce that the solution curve just discussed continues into that region and persists as $\sigma \rightarrow \infty$. For, the asymptote of $(O_1')_T$ is $\sigma = \frac{r_0^2}{u_1^2} \lambda$, and between it and $\lambda = 1$, which is the asymptote of $(I_B')_T$, $G(T(\lambda_1), \sigma)$ takes on all values between 0 and $-\infty$; in particular, the limited range of values assumed by $G(\lambda_1, \sigma)$ in this region. The behavior of the solution curve for large σ may be deduced by using the asymptotic formulae for curves $G(\lambda_1, \sigma) = g$ and $G(T(\lambda_1), \sigma) = g$ which are given in the appendix. These are

$$\sigma = -g\lambda_1 - gF\left(\frac{r_0}{\sqrt{-g}}\right),$$

and

$$\sigma = \frac{r_0^2}{u_1^2} \lambda_1 - \left[p + 2 \left(1 + \frac{gu_1^2}{r_0^2} \right) \frac{1 - \frac{r_0^2}{u_1^2}}{1 - u_1^2} \right].$$

It is clear that g at a point of intersection is given by $-r_0^2/u_1^2$ plus terms of order $1/\lambda_1$; substituting this value in the second equation gives the solution curve correctly to order $1/\lambda_1$ in the form.

$$\sigma = \frac{r_0^2}{u_1^2} \lambda_1 - p.$$

When the solution curve has such a linear asymptote it is convenient to calculate β^2 from the formula

$$\beta^2 = 1 + \frac{p}{\sigma} - \frac{\lambda}{\sigma} + \text{terms of order higher than } 1/\sigma$$

which is readily obtained from (30). In the present case

$$\beta^2 = \left(1 - \frac{u_1^2}{r_0^2} \right) \left(1 + \frac{p}{\sigma} \right) + \text{higher terms in } 1/\sigma. \quad (38)$$

As $\sigma \rightarrow \infty$, β^2 tends to the value appropriate to the TE_{11} -mode in an isotropic medium ($\mu \rightarrow \mu_z = \mu_0$, $\kappa \rightarrow 0$ as $\sigma \rightarrow \infty$). Thereby the whole solution curve is classified as specifying part of a TE_{11} -limit mode.

The remaining section of the TE_{11} -limit mode in the upper half-plane is again found in the region between $(O_1')_T$ and $(I_B')_T$ for $\sigma < \sigma_0$. Any line $\sigma = \text{constant} < \sigma_0$ cuts these two curves at two values of λ_1 . As λ_1 varies between these values, $G(T(\lambda_1), \sigma)$ varies from 0 to $-\infty$; it is, thus,

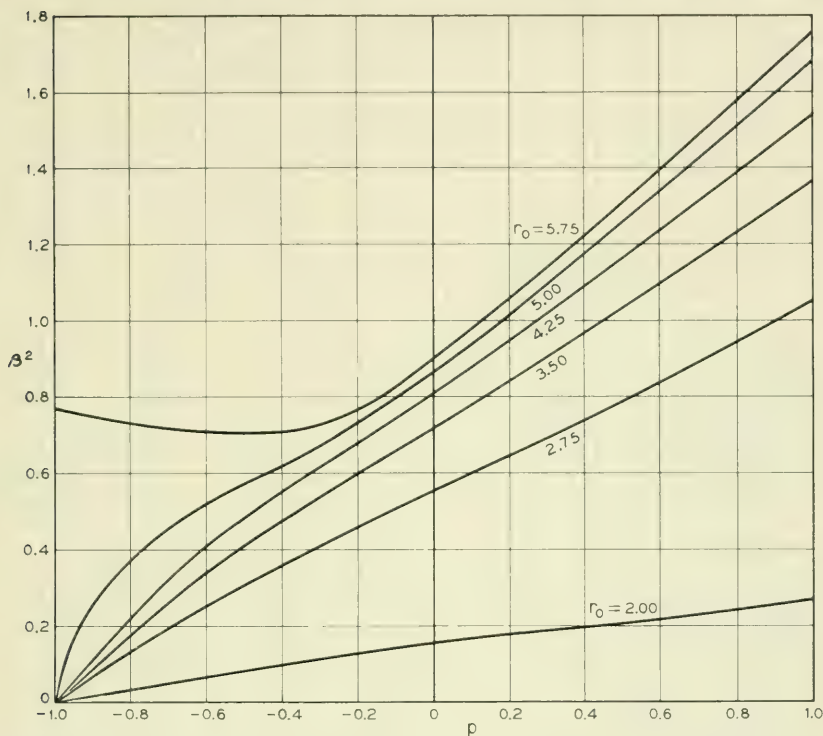


Fig. 10(a) — β^2 versus p for small values of σ — the TE_{11} -limit mode.

clearly equal to the finite (negative) $G(\lambda_1, \sigma)$ somewhere between. This situation persists up to $\sigma = \sigma_0 - 0$ and a solution curve therefore exists between $\sigma = 0$ and $\sigma = \sigma_0$. It meets $\sigma = 0$ for λ_1 satisfying

$$\frac{1}{1 - \lambda_1^2} \left[\frac{1}{\lambda_1} F(r_0 \sqrt{1 - \lambda_1^2}) - 1 \right] = \frac{1}{1 - \lambda_2^2} \left[\frac{1}{\lambda_2} F(r_0 \sqrt{1 - \lambda_2^2}) - 1 \right],$$

$$\text{and} \quad \lambda_1 + \lambda_2 = p. \quad (39)$$

These equations have been solved numerically; the corresponding $\beta^2 = -\lambda_1 \lambda_2$ is shown in Fig. 10(a). For r_0 between u_1 and j_1 a value derived for β^2 from the first three terms of an expansion of β^2 in powers of p , equation (61), turns out to be in very good agreement with the numerical calculation up to $p = 1$, for $\sigma = 0$ and presumably is good for small σ .

At σ_0 (the point at which μ becomes negative), the solution curve is "cutoff". However, the corresponding β^2 is not zero. As σ_0 is approached from below $G(\lambda_1, \sigma) \rightarrow 0$ and so $G(\lambda_2, \sigma)$ tends to zero. Thus, λ_2 tends to

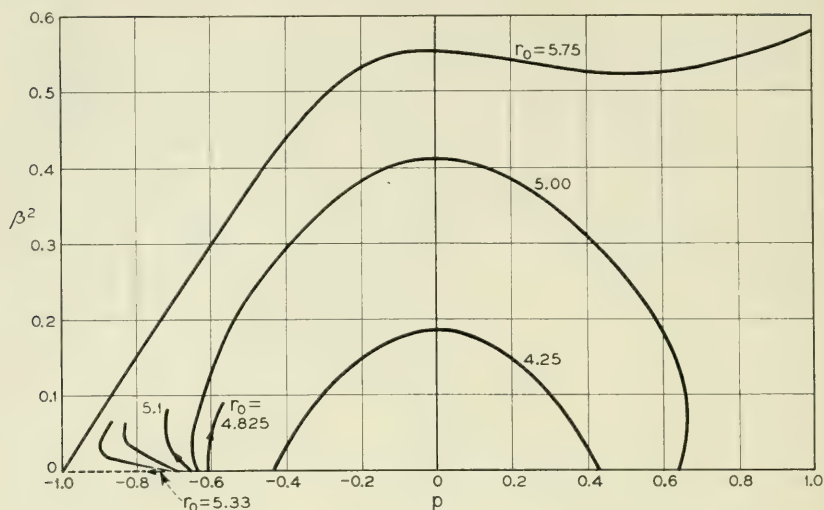


Fig. 10(b) — β^2 versus p for small values of σ — the TM_{11} -limit mode.

the negative root λ_{20} (unique for the present radius) of

$$F\left(r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma_0 \lambda_2}}\right) = \lambda_2.$$

The associated β^2 is $-\lambda_{20}\lambda_{10} = -\frac{\lambda_{20}}{\sigma_0}$ and is shown in Fig. 11(a). The way in which β^2 approaches this value as $\sigma \rightarrow \sigma_0$ can be found and is one of the more subtle examples of behavior of a mode near a special point. Writing $\sigma = \sigma_0 - \delta\sigma$, $\lambda_1 = \lambda_{10} - \delta\lambda_1$, we observe that, since $\sigma_0 + p - \frac{1}{\sigma_0} = 0$, the Polder relation in the form

$$\lambda_1 = \frac{1}{\sigma} + \frac{\sigma + p - 1/\sigma}{1 - \sigma\lambda_2}$$

fully determines $\frac{d\lambda_1}{d\sigma}$; any variation due to $\delta\lambda_2$ vanishes at $\sigma = \sigma_0$. $\delta\lambda_2$ can be determined from the G -equation. Near $\sigma\lambda = 1 - 0$ ($\lambda > 1$), $G(\lambda_1, \sigma)$ is given by

$$-\frac{r_0}{\lambda_1} \sqrt{\frac{1 - \sigma\lambda_1}{\lambda_1^2 - 1}},$$

which near σ_0 , λ_{10} may be written

$$-\sqrt{\delta\sigma} \frac{r_0}{\lambda_{10}} \sqrt{\frac{\lambda_{10} \frac{d\sigma}{d\lambda_1} + \sigma_0}{\lambda_{10}^2 - 1}}.$$

The perturbed $G(\lambda_2, \sigma)$ which (since $G(\lambda_{20}, \sigma_0) = 0$) is $\frac{\partial G}{\partial \lambda_{20}} \delta\lambda_2 + 0(\delta\sigma)$ equals the preceding expression and gives

$$\delta\lambda_2 = -\sqrt{\delta\sigma} \frac{r_0}{\lambda_{10}} \sqrt{\frac{\lambda_{10} \frac{\delta\sigma}{\delta\lambda_1} + \sigma_0}{\lambda_{10}^2 - 1}} \left(\frac{\partial G}{\partial \lambda_{20}} \right)^{-1}.$$

Accordingly, $\delta\beta^2 = -\lambda_{10} \delta\lambda_2 + 0(\delta\sigma)$, a result which shows that β^2 tends to its terminal value along the vertical. It is clear analytically and graphically that this mode persists as $p \rightarrow 0$, and must be identified with the only isotropic mode for this radius, namely TE_{11} . No other branches exist below $\sigma = \sigma_0$, since $G(\lambda_1, \sigma)$ and $G(T(\lambda_1), \sigma)$ have opposite signs except in the region just considered.

The two solution curves considered so far are not the only ones; in fact the infinity of sheets of the surface $G(\lambda_1, \sigma)$ in the region bounded by I_1, O_c and $(I_A')_T$, Fig. 9(a), intersect the transformed sheets $G(T(\lambda_1)\sigma)$ in infinitely many more curves. In the blank areas of that region the G -functions have equal sign, and all these areas must be carriers of solution curves, since in every one of them every single contour $G(\lambda_1, \sigma) = g$

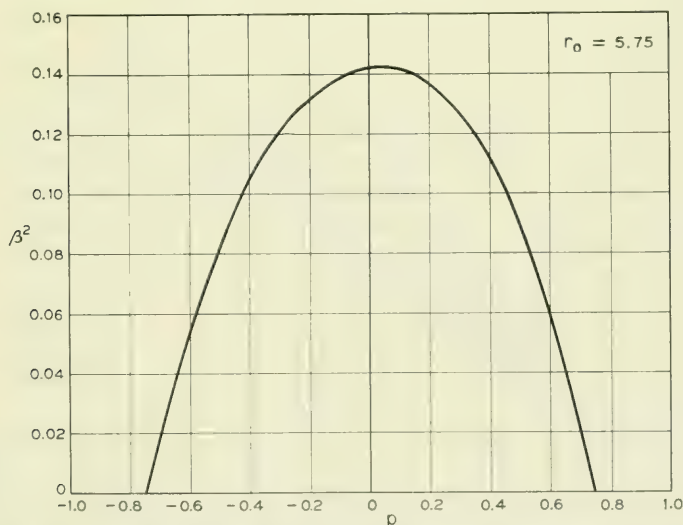


Fig. 10(c) — β^2 versus p for small values of σ — the TE_{12} -limit mode.

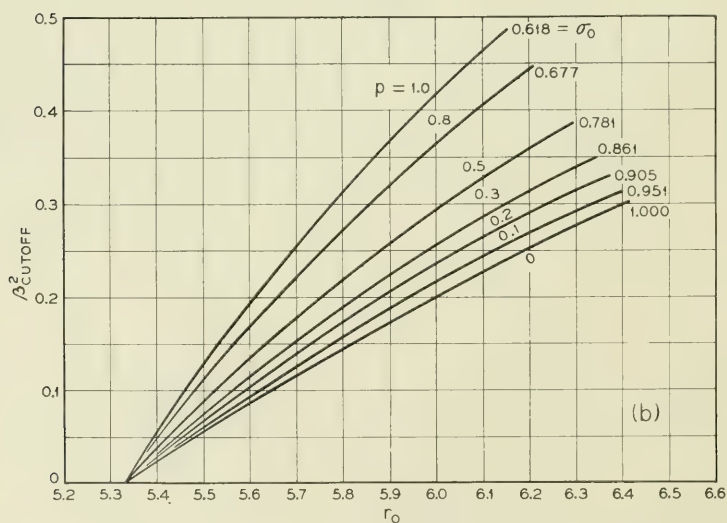
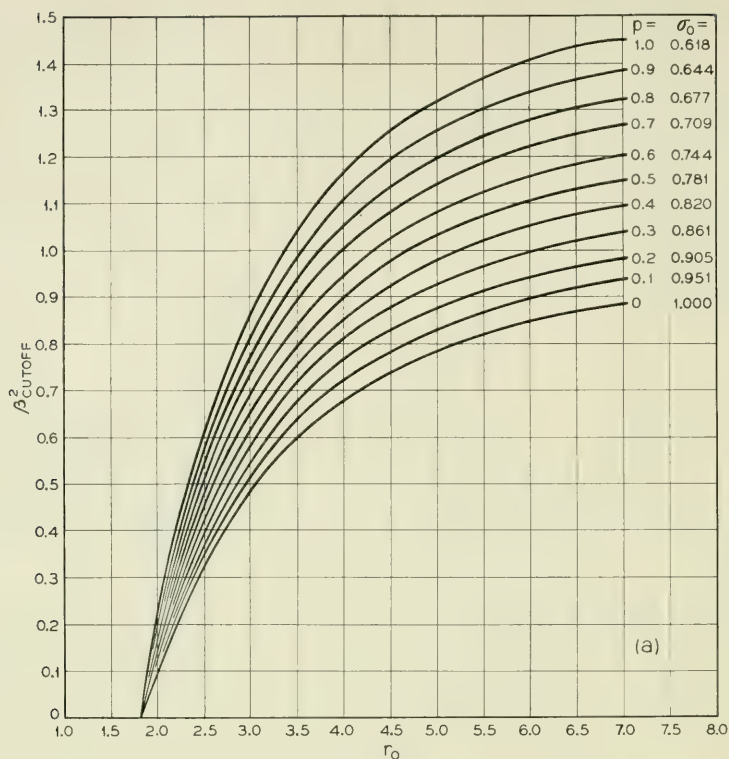


Fig. 11 — β^2 at the type $2'$ cutoff as a function of r_0 for various p . $2_0'$ cutoff, TE_{11} -limit mode; (b), $2_1'$ cutoff, TM_{11} -limit mode. The presence of a curve for $p = 0$ is clarified in the text.

crosses all contours of $G(T(\lambda_1), \sigma)$, in particular $G(T(\lambda_1), \sigma) = g$. All the additional solution curves arising in this way start at $\sigma = 1, \lambda_1 = 1$; the n^{th} of them threads its way from one blank region to another, first through the intersection of I_{n+1} with $(I_B')_T$, then through the intersection of 0_n with $(0_1')_T$, and finally comes to an end at the intersection of I_n with $(I_A')_T$. At the end point ($\sigma = 1, \lambda_1 = 1$), λ_2 and, therefore, β^2 are infinite, (just as for the TE_{11} solution curve). At the end point $(I_n, (I_A')_T)$, λ_2 , and, therefore, β^2 are zero. The σ and λ_1 values corresponding to the latter are obtained from the equations

$$\frac{1 - \lambda_{1n}^2}{1 - \sigma_n \lambda_{1n}} = \frac{j_n^2}{r_0^2}; \quad \lambda_{1n} = \sigma_n + p. \quad (40)$$

It is possible to derive the slope $\delta\beta^2/\delta\sigma$ of the $\beta^2 - \sigma$ curves at these cut-off points. Near cut-off, the infinity I_n of $G(\lambda_1, \sigma)$ is matched by the infinity I_A' of $G(\lambda_2, \sigma)$. The G -equation therefore degenerates to

$$\frac{1}{\lambda_2} F(r_0) = \frac{j_n}{r_0 \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma \lambda_1} - j_n}} \cdot \frac{1}{\lambda_{1n}} \cdot \frac{r_0^2}{j_n^2}.$$

Writing $\sigma = \sigma_n - x\lambda_2$, $\lambda_1 = \lambda_{1n} - y\lambda_2$, expansion of the right hand side of this equation to order $1/\lambda_2$ furnishes one relation between x and y ; the Polder equation furnishes another. The two can be solved for x , and so, since to first order

$$\delta\beta^2 = -\lambda_{1n}\lambda_2 = \lambda_{1n} \frac{\sigma - \sigma_n}{x} = \lambda_1 \frac{\delta\sigma}{x}$$

$\delta\beta^2/\delta\sigma$ may be found. It is found that for convenience in computation, the results of this calculation are best presented parametrically. Equations (46-8) represent equations (40) and $\delta\beta^2/\delta\sigma$ in this way. Fig. 12 (a) and (b) show the result of some computations. Near $\sigma = 1$, $\beta^2 = \infty$, these added solution curves behave rather like the TE_{11} curve. The leading term in the expansion of β^2 in powers of $\frac{1}{\sigma - 1}$ is now

$$\frac{p \left(1 - \frac{z_{n+1}^2}{r_0^2} \right)}{\sigma - 1}$$

for the solution curve ending at $I_n - (I_A')_T$. Here z_{n+1} is the $(n+1)^{\text{th}}$ root of $F(z) = 1$, not counting 0.

It will turn out later that the infinity of solution curves just discussed represents an incipient form of the whole mode spectrum; the reservoir

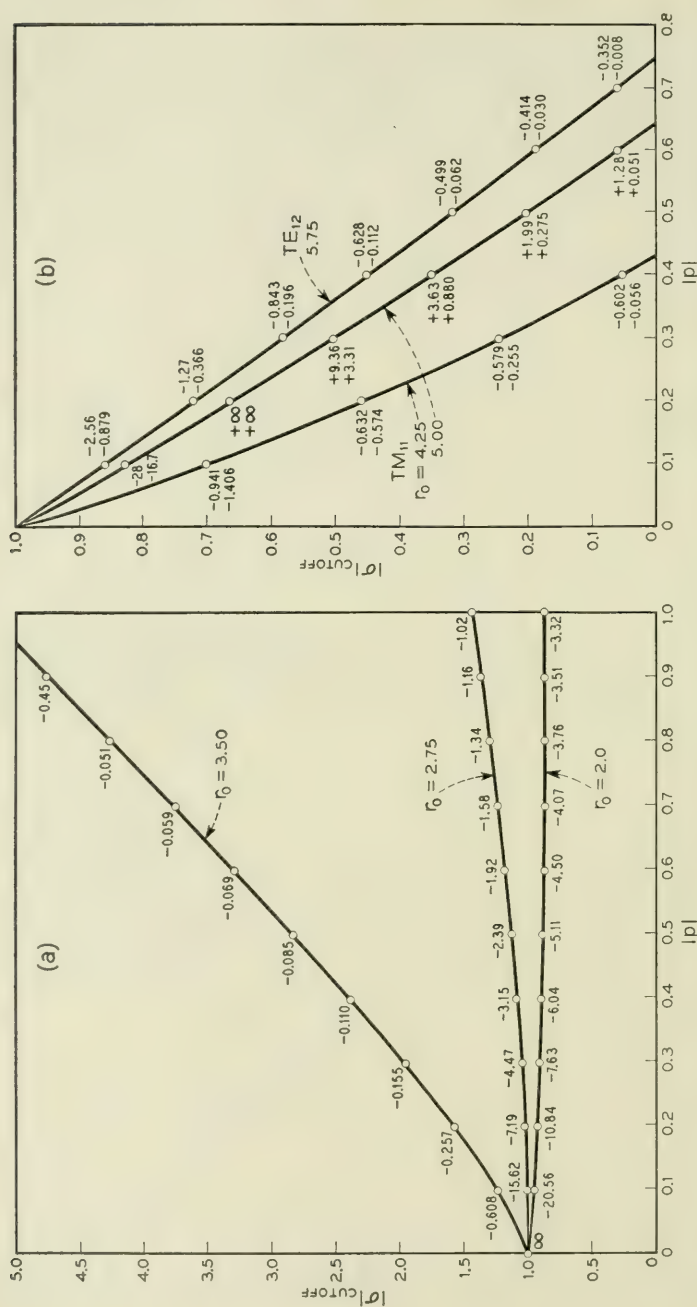


Fig. 12—Type 1 cutoff. $|\sigma|$ versus $|p|$ for some values of r_0 . (a) refers to the TM_{11} -limit mode in its incipient stage; the number attached to the points are values of $\frac{d\beta^2}{d\sigma} \text{sgn } p$ at the cutoff. (b) refers to the fully developed TM_{11} -limit and TE_{12} -limit modes in the region $|\sigma| < |\sigma_0|$; the upper attached number is $\frac{d\beta^2}{d\sigma} \text{sgn } p$ and the lower attached number is $\frac{d\beta^2}{dp} \text{sgn } p$.

from which higher modes are drawn as the guide radius is increased. That the propagation of modes which for larger guide radii correspond to higher TM and TE modes is possible for limited ranges of σ might be ascribed to the larger μ -values in those ranges, which cause the wave to see an effectively larger guide. This explanation is convincing only when $\sigma > 1$. When $\sigma_0 < \sigma < 1$, μ is negative, and the propagation must then be the result of an interplay between μ and κ . In passing we remark that we are here dealing with the propagation analog of so-called "shape resonances," which physicists sometimes encounter in resonance experiments on small spheres of ferrite in cavities.

We now turn to a discussion of the solution curves for $\sigma < 0$ which lie in the third quadrant. Fig. 9(b) shows the partition of the region allowed by the Polder relation (again for $p = \frac{3}{8}$) into positive and negative regions by the various I' , $0'$, $(I)_T$ and $(0)_T$ curves. Regions in which $G(\lambda_2, \sigma)$ and $G(T(\lambda_2), \sigma)$ have opposite signs are shaded. For $\sigma < -\sigma_0$, the question whether a given region of like signs is the site of a solution curve may, with one exception, be answered by the same type of geometrical argument as used for $\sigma > 0$. The singular area is that part of the region bounded by I_B' and 0_c in which the G -functions are both positive. Here both $G(\lambda_2, \sigma)$ and $G(T(\lambda_2), \sigma)$ are zero on 0_c ; $G(\lambda_2, \sigma)$ goes to ∞ on I_B' , whereas $G(T(\lambda_2), \sigma)$ is finite throughout the region. No intersection can be predicted, then, by the earlier argument. It can indeed be shown (for all r_0) that there is no such intersection. For, in the case $p = 0$, the solution curves are I_n' or $0_n'$ curves as demonstrated in Section (4.11). The region under consideration contains no such curves, and hence no solution curves. Thus, for $p = 0$, since $G(\lambda_2, \sigma)$ goes to infinity on I_B' , the surface $G(T(\lambda_2), \sigma)$ must lie entirely below the surface $G(\lambda_2, \sigma)$. Consider now, for fixed σ and increasing p , a point on the $G(T(\lambda_2), \sigma)$ surface whose height remains unchanged. For such a point $T(\lambda_2, p, \sigma)$ remains fixed and from the Polder relation this means an increasingly negative λ_2 . Since it can be shown that $G(\lambda_2, \sigma)$ goes monotonically from 0 to ∞ as λ_2 becomes more negative, it follows that $G(T(\lambda_2), \sigma)$ continues to be below $G(\lambda_2, \sigma)$ for all p .

All other regions of common sign do carry solution curves. That corresponding to the TE₁₁-limit mode begins at $\sigma = -1$, $\lambda_2 = -1$, passes through the intersection of $(0_0)_T$ and $0_0'$ and persists for indefinitely large σ . The asymptotic formula (38), for β^2 at large σ also holds as $\sigma \rightarrow -\infty$, if the signs of both σ and p are taken to be negative. The behavior of β^2 near $\sigma = -1$ may be found by the same means used at $\sigma = 1$. The resulting expression* is to order $\frac{1}{\sigma + 1}$, essentially the same

* See Section 4.17 for a more exact formula.

as the earlier (37) except that the smallest root of $F(z) = -1$ replaces that of $F(z) = 1$. The remaining solution curves confined to the region bounded by $0_0'$, 0_c , $(I_A)_T$ portray the incipient modes already encountered in the first quadrant. Their behavior near $\sigma = -1$ also follows (58), associated with the higher roots of $F(z) = -1$. Their end points, the intersections of I_n' with $(I_A)_T$, are still given by the parametric representation (46-8), due regard being paid to the signs of σ_n and p .

The remaining branch of the TE_{II}-limit mode, lying above $\sigma = -\sigma_0$, is found in the triangle between $(I_A)_T$, $\sigma = 0$ and I_B . Its end points are given by (39) with p negative and by the intersection of I_B' with $(I_n)_T$ which is $\sigma = -(1 + p)$, $\lambda_2 = -1$, $\lambda_1 = 0$. Thus the cut-off in contrast to the analogous branch for $\sigma > 0$, is given by $\beta^2 = 0$, $\sigma = -(1 + p)$. (When $p < -1$, the branch does not exist at all.) We note that a left-circular plane wave is cut off at exactly the same value of σ as the TE-mode is in this particular case (see, however, the following sections). The slope at cut-off is determined by expanding the G functions near their infinities at I_n and I_B' and utilizing the Polder relation. The slope is found to be

$$\frac{d\beta^2}{d\sigma} = \frac{F(r_0)}{p(1 - F(r_0))}. \quad (41)$$

A further solution curve lies in the region between 0_c and $(0_0)_T$ for $\sigma > -\sigma_0$. It has no analogue in a guide with isotropic material and will be discussed later.

In the discussion of the mode spectrum for radii between u_1 and j_1 three distinct types of cut-off point have already been encountered. When larger radii are treated it is found that no other types arise.* In Section 4.17 formulas relevant to the three types are given. An examination of the field components in the neighborhood of the cut-off points is of some interest. Cut-off points of type one (intersections of I_n and $(I_A')_T$ or I_n' and $(I_A)_T$), at which $\beta^2 = 0$, have $E_z = 0$ and the field is of a pure TE-type. The medium behaves transversely as though it had a permeability, $\mu = \kappa^2/\mu$. Although the field is purely TE at cut-off the mode terminating at such a point may in the limit of vanishing magnetization be either a TE- or a TM-mode. This impartiality extends to cut-off points of the other types. Cut-off points of type II $[(0_n')_T - 0_c$ or $(0_n)_T - 0_c]$ occur at $\sigma = \pm\sigma_0$, where $\mu = 0$ and here β^2 does not vanish. In such cases one of the χ 's is finite and the corresponding contributions to the field pattern quite normal. The other, however, tends

* There is an exception to this statement. This is the type designated in Section 4.17 as $2_0\infty$ which cuts off an isolated mode having no TE or TM analogue.

to an infinite imaginary value and the associated fields are confined very closely to the guide walls. The wall currents are very large and essentially longitudinal. Type III cut-off points $[I_H - (I_A)_T]$ at which $\mu = -\kappa$ have $\beta^2 = 0$, but the fields are not of a purely TE- or TM-type. They consist essentially of a rotating, transverse, H , which is uniform over the guide. The components H_z , E_θ and E_r are smaller by one order of $\sigma - \sigma_{\text{cut-off}}$ and E_z , two orders smaller.

It should be stressed again that, in general, the modes are never of pure TE or TM type. Nevertheless, for the sake of brevity, we shall refer to them as such; calling them TE-modes or TM-modes according to their limit as the magnetization is removed.

4.13. We now consider the behavior of the modes as a function of radius. The reader will be aided by Figs. 8(a) to (e) and 9(a) to (g). In preparation for this it is necessary to examine the movement of the I_n , I_n' and 0_n , $0_n'$ curves when r_0 is varied. It will be recalled that the equation for the I_n curves and their reflections, I_n' , in the origin, is

$$\frac{1 - \lambda^2}{1 - \sigma\lambda} = \frac{j_n^2}{r_0^2}.$$

The contours $\chi(\lambda, \sigma) = C$, where $\chi^2 = (1 - \lambda^2)/(1 - \sigma\lambda)$ have already been plotted in Fig. 4. The I_n , I_n' curves are among these, and, clearly, for a fixed n , the associated χ^2 decreases as r_0 increases. The course of a given pair (I_n , I_n') may then be seen directly from Fig. 4. The qualitative behavior of the pair changes radically only when r_0 passes through the value j_n . Before it does so, I_n lies, for λ between 0 and 1, above $\sigma = \lambda$ and tends to $\sigma = \infty$ as λ tends to zero. At $r_0 = j_n$, the I_n and I_n' curves merge into the lines $\sigma = \lambda$ and $\lambda = 0$. Beyond j_n , I_n lies below $\sigma = \lambda$ for λ between 0 and 1 and goes to $-\infty$ as λ approaches zero. The I_n' -curve remains, throughout the reflection of I_n in the origin. As $r_0 \rightarrow \infty$, I_n tends to the line $\lambda = 1$, I_n' to $\lambda = -1$. No I_n curves ever enter the region $\lambda > 1$, $\sigma < 0$; no I_n' curves enter $\lambda < -1$, $\sigma > 0$. It is also important to relate the I_n , I_n' curves to the boundaries of the Polder regions. I_n curves cut the Polder boundary $\sigma = \lambda - p$, of the first quadrant in at most one point. As r_0 increases from 0 to j_n , this point moves from $\sigma = \sigma_0$ to $\sigma = \infty$. Thereafter, no intersection occurs at fixed p until r_0 equals $j_n/\sqrt{1 - p^2}$; it here reappears at $\sigma = 0$ and moves steadily to $\sigma = 1 - p$ as r_0 increases indefinitely. The only intersection with the other Polder boundary $\sigma = 1/\lambda$, is at $\lambda = 1$, $\sigma = 1$, regardless of r_0 .

The 0_n , $0_n'$ curves are given by

$$\sigma = \frac{1}{\lambda} \left[1 - r_0^2 \frac{1 - \lambda^2}{(F^{-1}(\lambda))^2} \right]$$

if the n^{th} branch of $F^{-1}(\lambda)$ is used. Thus, as r_0 increases, the successive curves either all pass through a fixed point (which can only be $\lambda = \pm 1$, $\sigma = \pm 1$, $n > 0$) or move steadily up or down without further intersection. An 0_n curve starts from $\sigma\lambda = 1$ at $r_0 = 0$ and falls for $\lambda < 1$, rises for $\lambda > 1$, as r_0 increases. For large λ , since $\sigma \sim j_n^2/r_0^2 (\lambda - 2)$, $n > 0$, the 0_n and I_n curves move together with a constant separation. 0_0 is singular, since it does not pass through λ , $\sigma = 1$ and falls steadily for all λ ; it tends to $\sigma = 0$ for large λ . The $0_n'$ curves rise from $\sigma\lambda = 1$ at $r_0 = 0$ for $-1 < \lambda$, fall for $\lambda < -1$. They run parallel to I_{n+1}' for $-\lambda$ very large. For small λ there is an expansion

$$\sigma = \left(1 - \frac{r_0^2}{u_n^2}\right) \frac{1}{\lambda} - \frac{2r_0^2}{u_n^2(u_n^2 - 1)} + O(\lambda)$$

holding for 0_n and for $0_n'$. This indicates that for $r_0 < u_n$, 0_n goes to $+\infty$ and $0_n'$ to $-\infty$ for small λ , but at $r_0 = u_{n+1}$, 0_n and $0_n'$ merge momentarily at

$$\lambda = 0,$$

$$\sigma = -\frac{2r_0^2}{u_{n+1}^2(u_{n+1}^2 - 1)} < 0.$$

For larger r_0 , 0_n goes to $-\infty$ and $0_n'$ to $+\infty$. Since the union of 0_n and $0_n'$ takes place at a negative σ , it is clear that 0_n curves, unlike I_n curves, may cross the line $\sigma = 0$ twice. Intersections of the 0_n , $0_n'$ curves with the Polder boundary are difficult to examine explicitly and this may lead to some obscure situations for $0 < |\lambda| < 1$. However, for $\sigma > \sigma_0$, since 0_n and I_n have a fixed separation for large $|\lambda|$, this pair escape intersection with the boundary at the same value of r_0 , namely j_n . Similarly $0_n'$ and I_{n+1} escape together at $r_0 = j_{n+1}$ for $\sigma < -\sigma_0$.

We shall now examine the effect of varying r_0 upon the sequence of modes when $\sigma > \sigma_0$. When r_0 is less than u_1 , a case in which the isotropic medium would not propagate, no part of $0_0'$ lies in the upper half plane and there is then no $(0_0')_T$ curve. The solution curve which in the previous discussion of Section 4.12 was assigned to TE_{11} , after passing the intersection $[I_1 - (I_B)_T]$ can no longer escape to infinity and terminates on $[I_1 - (I_A)_T]$. Thus, the TE_{11} mode at this radius has become an incipient mode with cut-off and other properties given by the formulae already quoted for such modes. As r_0 approaches u_1 from below, the $\beta^2 - \sigma$ curve is double valued between $\sigma_{\text{cut-off}}$ and some larger value. This is borne out by the fact that $d\beta^2/d\sigma$ becomes positive at cut-off, and by the observation that the solution curve bulges towards large σ between $I_1 - (I_B')_T$ and its terminus. The part of the $\beta^2 - \sigma$

curve along which $\frac{d\beta^2}{d\sigma} < 0$, will tend smoothly towards the $\beta^2 - \sigma$ curve for r_0 just greater than u_1 . The course of the TE_{11} solution curve remains qualitatively unchanged for all $r_0 > u_1$.

When r_0 passes through u_1 , and the TE_{11} solution curves escapes discontinuously to infinity, the solution curves below it disengage from their former end points $I_{n+1} - (I_A)_T$ and instead end at the point $I_n - (I_A)_T$. When r_0 exceeds j_1 , the curves I_1 and 0_1 escape intersection with $\sigma = \lambda - p$ simultaneously, for $\sigma > \sigma_0$, and the curve $(I_1')_T$ makes its first appearance. From the asymptotic formulae (App. II) the latter runs to infinity between I_1 and 0_1 , and now the solution curve which ended for $u_1 < r_0 < j_1$ at $I_1 - (I_A)_T$ is carried to infinity between I_1 and $(I_1)_T$. The asymptotic expression for β^2 versus σ , given in formula (56) indicates that β^2 tends to the isotropic value for the TM_{11} mode. No further qualitative changes will take place in behavior of this mode as r_0 increases.

As r_0 increases through u_2 (the value at which the isotropic medium supports the TE_{12} mode), the $(0_1')_T$ curve makes its appearance, an event accompanied by the escape of the uppermost incipient solution curve (the one ending at $(I_A)_T - I_2$) to infinity. The escape takes place in the same way as that of the TE_{11} solution curve as r_0 passed through u_1 . The newly escaped curve, of course, represents the TE_{12} -limit mode. The end points of the remaining incipient solution curves also jump discontinuously to their next higher neighbors as they did at $r_0 = u_1$. The course of events as r_0 is increased further should now be abundantly clear, and is summarized in Table I on page 642.

We now turn to the region $0 < \sigma < \sigma_0$ and consider first the situation $0 < r_0 < u_1$. It is clear that in the area bounded by 0_∞ , 0_0 , $(I_A)_T$ and $(I'_B)_T$ both G functions are negative. There is no simple geometrical argument which determines the existence of a solution curve in this region. It is therefore necessary to use a type of analytic argument, which is useful in a number of other cases, although fully discussed only in the present instance.

We show that the least value attained by $G(\lambda_1, \sigma)$ in the admissible region for $p = 0$ (which contains all regions admissible for other p -values) is greater than the maximum value of $G(\lambda_2, \sigma)$ in the range $-1 < \lambda_2 < 0$, $\sigma > 0$. Consider the variation of $G(\lambda_1, \sigma)$ as the point $\lambda_1 = 1$, $\sigma = 1$ is approached along a line of constant χ^2 in the admissible region for $p = 0$ (see Fig. 4). We have the relation

$$G(\lambda, \sigma) = \frac{1}{\chi^2} \left[\frac{1}{\lambda} F(r_0 \chi) - 1 \right].$$

For χ^2 negative, $F(r_0\chi)$ is positive and thus, as λ approaches unity from above G decreases. Again for χ^2 positive and $r_0 < u_1$, $F(r_0\chi)$ is positive and thus, as λ approaches unity from below, G also decreases. Thus for any χ^2 , $G(\lambda, \sigma)$ takes on its least value at $\sigma = 1$, $\lambda = 1$ and this value is

$$\frac{1}{\chi^2} [F(r_0\chi) - 1].$$

The minimum value of this limit in this region is $F(r_0) - 1$ and is greater than -1 . In the region $-1 < \lambda_2 < 0$, $\sigma > 0$, χ^2 is between 0 and 1, and the χ^2 curves run from $\sigma = 0$ to $\sigma = \infty$. G will clearly decrease as σ increases from zero on any one of these curves. Thus G attains its maximum on $\sigma = 0$, where its value is

$$- \frac{1}{1 - \lambda^2} \left[\frac{F(r_0 \sqrt{1 - \lambda^2})}{|\lambda|} + 1 \right].$$

Since $F(r_0 \sqrt{1 - \lambda^2})$ is positive for $r_0 < u_1$ and $|\lambda| < 1$, G is clearly less than -1 . In passing we note that for $r_0 = u_1$, both G functions may attain the value -1 .

As r_0 passes through u_1 , the $(0_0')_T$ curve appears in the region under discussion and together with $(I_B')_T$ delimits the region carrying the TE_{11} -solution curve already discussed at length. No qualitative changes occur in that curve as r_0 is increased indefinitely. When r_0 exceeds j_1 , the $(I_1')_T$ curve appears between $(0_0')_T$ and $(I_A')_T$. Between $(I_A')_T$ and $(I_1')_T$ the G functions have a region of common sign, yet no solution curve arises there for a given p until r_0 reaches $j_1/\sqrt{1 - p^2}$.* From then on, the I_1 curve cuts $(I_A')_T$, see Fig. 9(c), and a solution curve exists between $(I_1')_T$ and I_1 . It is cut off at the intersection $(I_A')_T - I_1$; there, $\beta^2 = 0$ and $\sigma, \frac{d\beta^2}{d\sigma}$ are given by the same parametric formulae (46-8)

applying to the cut-off of incipient modes, the parameter θ being negative. The curve begins at $\sigma = 0$, where it satisfies the usual equation, which for this radius has two solutions. The solution with the smaller λ , belonging to the present curve, tends to the isotropic TM_{11} -limit as $p \rightarrow 0$. At a fixed r_0 , sufficiently below u_2 , this mode does not exist at

* There are some exceptions to this statement. When $4.82 < r_0 < u_2 = 5.33$ and p exceeds $\sqrt{1 - j_1^2/r_0^2}$, a double-valued $\beta^2 - \sigma$ curve exists between two positive σ values. For values of r_0 still closer to u_2 further regions of common sign may arise as a result of the interplay of the $(0_1')_T$ and $(I_A')_T$ curves. We have not examined these regions closely. Such dubious regions are confined to the immediate neighborhoods below the u_n .

all when

$$p > \sqrt{1 - \frac{j_1^2}{r_0^2}}.$$

If r_0 is greater than u_2 , the $(0_1')_T$ curve has appeared. A new region of like signs of the G 's arises between it and $(I_1')_T$, see Fig. 9(e), and contains a solution curve. This ends at σ_0, λ_{10} and begins at $\sigma = 0$ at a value of λ_1 pertaining to the TM_{11} -mode. Thus, it is clear that as r_0 passed u_2 , the end-point of the TM_{11} curve jumped discontinuously from $(I_A')_T - I_1$ to σ_0, λ_{10} . This jump is anticipated as r_0 approaches u_2 ; the $\beta^2 - \sigma$ curve first bulges beyond $(I_A')_T - I_1$ towards its later course and returns to that point with positive slope. As r_0 increases further no change occurs in the qualitative behavior of the mode. It may be noted that above u_2 the mode exists for all p .

Beyond $r_0 = u_2$, at least part of the area between I_1 and $(I_A')_T$ is an admissible region and does in fact contain the TE_{12} solution curve. It begins at $\sigma = 0$ and λ_1 given by that solution of eqn. (39) which is, in the limit $p = 0$, the TE_{12} solution. It is cut off with $\beta^2 = 0$ at $(I_A')_T - I_1$, the end point relinquished by the TM_{11} -solution curve. As r_0 passes j_2 , the TE_{12} solution retains its cut-off point, but, beyond $r_0 = u_3$, it will transfer this point discontinuously to σ_0, λ_{10} . Thereafter its course remains essentially unaltered. Tables I, II and III show the progression of cut-off points of the various modes.

It may be recalled that in the analysis of $u_1 < r_0 < j_1$, the modes in $\sigma < -\sigma_0$ followed essentially the same course as in $\sigma > \sigma_0$. This is also true of their progress with changing radius and of the escape process. The singular character of the $(0_0)_T$ curve and the presence of I_B lead to some local changes in the progress of the modes but have no effect on their more salient features in this particular range of σ . The scheme of progression of the end points is shown in Table I.

In contrast with the state of affairs in the region just discussed, the mode structure in the area between $\sigma = 0$ and $\sigma = -\sigma_0$ is very markedly affected by the presence of $(0_0')_T$ and I_B' .

When $r_0 < u_1$, a solution curve exists between $\sigma = -1 - p$, and $\sigma = -\sigma_0$. It starts with $\beta^2 = 0$ at the intersection of $(I_A)_T$ and (I_B') with a slope given by (41). For sufficiently small r_0 , β^2 tends to infinity as $\sigma \rightarrow \sigma_0$, since the solution curve approaches the line $0_c'$ or $(0_x)_T$. Its shape is then given by (52), see Section 4.17. As r_0 increases, 0_0 falls steadily. Eventually, for sufficiently large p , its minimum falls below

* See footnote on page 628.

$-\sigma_0$. $(0_0)_T$ now has two branches for $\sigma > -\sigma_0$, which pass through the point $\sigma = -\sigma_0$, $\lambda_2 = -\frac{1}{\sigma_0}$ making there a finite angle with each other. $(0_0)_T$ is completed by a loop in $\sigma < -\sigma_0$, Fig. 9(g), which does not affect the incipient modes appreciably. The mode in question now has two branches. The first starts as before and ends at $\sigma = -\sigma_0$, where the associated β^2 is given by $\beta_a^2 = \lambda_a/\sigma_0$ and λ_a is the smaller root of

$$F\left(r_0 \sqrt{\frac{1 - \lambda_a^2}{1 - \sigma \lambda_a}}\right) = \lambda_a. \quad (42)$$

It resumes at $\sigma = -\sigma_0$ and $\beta^2 = \beta_b^2 = \lambda_b/\sigma_0$, where λ_b is the larger root of the above equation, progresses to smaller $|\sigma|$ -values and then back to $\sigma = -\sigma_0$ where β^2 tends to infinity again in accordance with (52). Beyond $r_0 = u_1$, where 0_0 rises steadily from $\sigma = -\infty$ to $\sigma = 0$ with increasing λ , one branch of $(0_0)_T$ in $-\sigma_0 < \sigma < 0$ disappears and only the second branch of the mode remains. Neither branch has an analogue in ordinary waveguides; as $p \rightarrow 0$ each lies in a smaller and smaller neighborhood of $\sigma = 1$, and finally vanishes into $\sigma = 1$, $\lambda_2 = -1$.

For r_0 between u_1 and j_1 there is a single solution curve starting at $\sigma = 0$ and ending with $\beta^2 = 0$ at $(I_A)_T - I_B'$. This may be identified in the limit $p = 0$, with the TE_{11} -limit mode, and has already been fully discussed for $u_1 < r_0 < j_1$. No change in the formula for its cut-off point occurs up to $r_0 = u_2$. A useful spot-point $(I_B' - (I_1)_T)$ along its course can be found when

$$0 < p < 1 - \sqrt{1 - \frac{j_1^2}{r_0^2}}$$

and is given by (60).

In the range $j_1 < r_0 < u_2$, a further solution curve (corresponding to the TM_{11} -limit mode) can arise, provided

$$p < \sqrt{1 - \frac{j_1^2}{u_2^2}}$$

The radius at which it will then first appear is

$$r_0 = \frac{j_1}{\sqrt{1 - p^2}}.$$

It begins on $\sigma = 0$, according to (39) and is cut off, with $\beta^2 = 0$, at $(I_A)_T - I_1'$.

As r_0 passes u_2 , the cut-off point of the TE_1 solution curve moves dis-

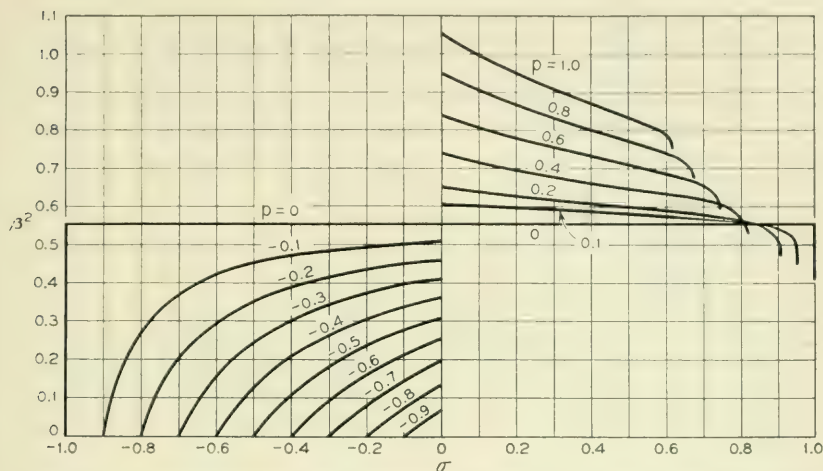


Fig. 13 — Approximate course of modes in the $\beta^2 - \sigma$ plane for various p values and at two values of r_0 . Fig. 13(a), above, $r_0 = 2.75$, TE_{11} -limit modes; $-1 < \sigma < 1$. Fig. 13(b), $r_0 = 2.75$, TE_{11} -limit mode and incipient TM_{11} -limit mode, $\sigma > 1$. Fig. 13(c), $r_0 = 2.75$, TE_{11} -limit mode, $\sigma < -1$. Fig. 13(d), (e) and (f), $r_0 = 5.75$, TE_{11} -limit and TE_{12} -limit modes; Fig. 13(d), $-1 < \sigma < 1$; Fig. 13(e), $\sigma > 1$; Fig. 13(f), $\sigma < -1$. Fig. 13(g), 13(h) and 13(i), $r_0 = 5.75$, TM_{11} -limit modes; Fig. 13(g); $-1 < \sigma < 1$; Fig. 13(h), $\sigma > 1$; Fig. 13(i), $\sigma < 1$. It should be noted that a scale linear in $\frac{\beta^2}{1 + \beta^2}$ is used for convenience when $|\sigma| > \sigma_0$.

continuously to $\sigma = -\sigma_0$, $\lambda_2 = -\frac{1}{\sigma_0}$ and, simultaneously, the cut-off point of the TM_{11} curve occupies the position relinquished by the former. The TM_{11} mode now exists for all p . A new solution curve (TE_{12}) appears in the region bounded by $0_1'$, $(0_1)_T$ and I_1' , if p is not too large, terminating at $(I_A)_T - I_1'$, the point left by the TM_1 terminus. (If p exceeds $\sqrt{1 - j_1^2/u_3^2}$ this curve will not exist at all.)

Figs. 13(a) to (i) and 14 show the approximate course of the $\beta^2 - \sigma$ curves for the TE_{11} mode at $r_0 = 2.75$ and for the TE_{11} , TM_{11} and TE_{12} modes at $r_0 = 5.75$. The incipient TM_{11} mode at $r_0 = 2.75$ is shown for positive σ , p only. They were computed by the methods outlined above.

4.14. *Guides of large radius.* It is of some interest because of the high dielectric constant of ferrites to examine the behavior of the modes as the radius, r_0 , is allowed to become very large. The two sides of the G -equation will remain determinate for unlimited r_0 provided

$$r_0 \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} \left(\text{or } r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \right)$$

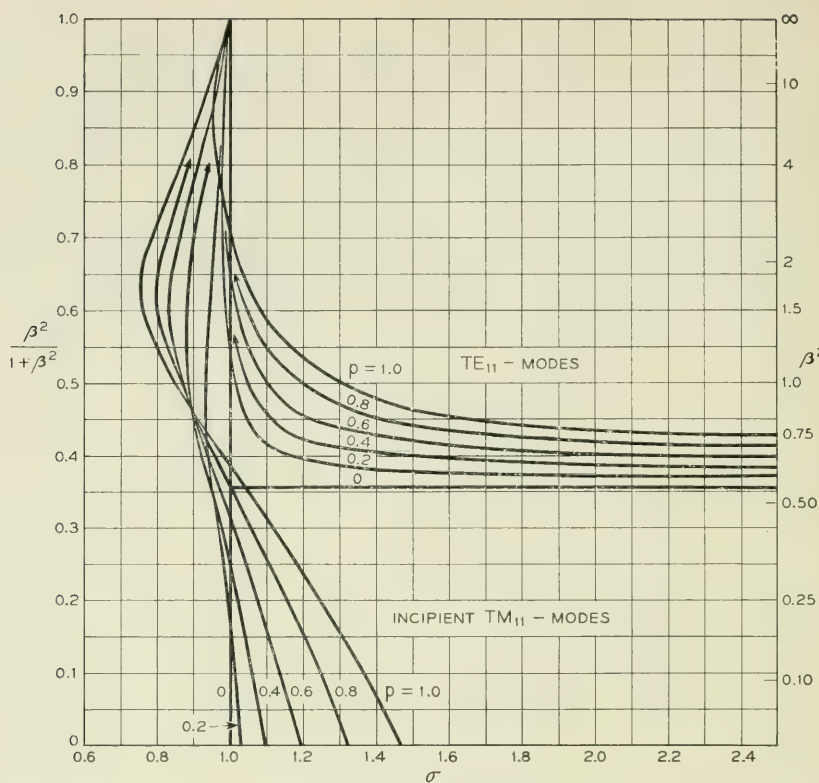


Fig. 13(b) — See Fig. 13.

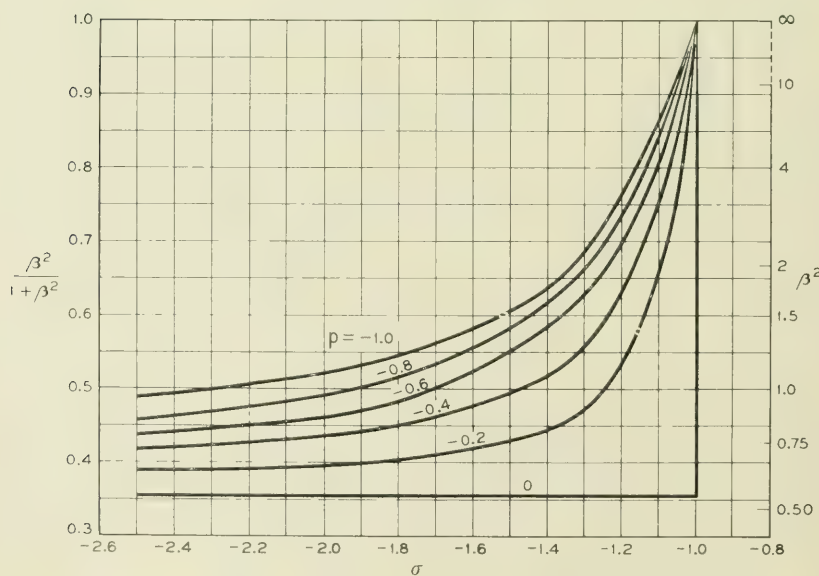


Fig. 13(c) — See Fig. 13.

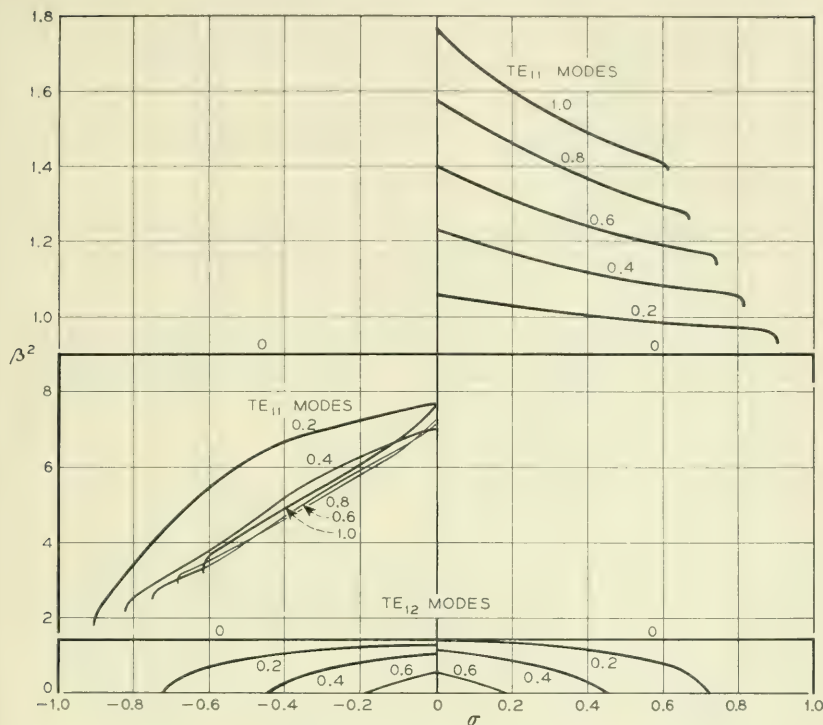


Fig. 13(d) — See Fig. 13.

remains finite, while

$$r_0 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma \lambda_2}} \quad \left(\text{or } r_0 \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma \lambda_1}} \right)$$

becomes infinite imaginary. Examining the solutions obtained under these conditions it is possible to find expansions for β^2 in inverse powers of r_0^2 . These are as follows:

for $p > 0$, $\sigma > 1$ or $p < 0$, $-\sigma_0 < \sigma < 0$

$$\beta^2 = 1 + \frac{p}{\sigma - 1} - \left(1 + \frac{p/2}{\sigma - 1} \right) \frac{x_n^2}{r_0^2}, \quad (43a)$$

where the x_n are the successive roots greater than zero of $F(x_n) = 1$ and the modes are associated with the x_n by the scheme: $\text{TE}_{11} \rightarrow x_1$, $\text{TM}_{11} \rightarrow x_2$, $\text{TE}_{12} \rightarrow x_3$, etc:

for $p > 0$, $0 < \sigma < \sigma_0$ or $p < 0$, $\sigma < -1$

$$\beta^2 = 1 + \frac{p}{\sigma - 1} - \left(1 + \frac{p/2}{\sigma + 1} \right) \frac{y_n^2}{r_0^2}, \quad (43b)$$

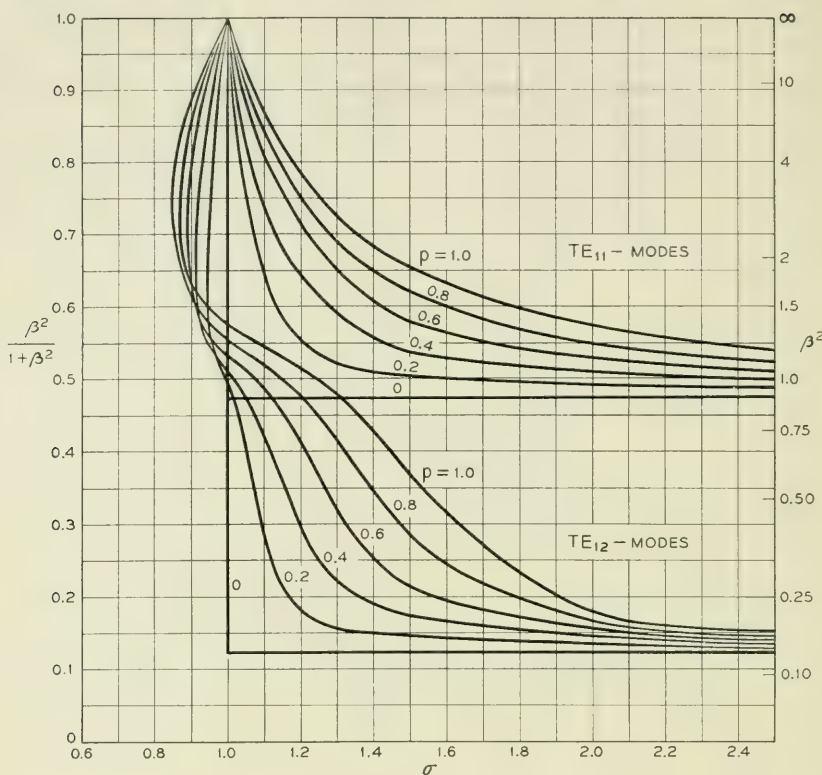


Fig. 13(e) — See Fig. 13.

with $F(y_n) = -1$ and $\text{TE}_{11} \rightarrow y_1$; $\text{TM}_{11} \rightarrow y_2$; $\text{TE}_{12} \rightarrow y_3$, etc.

for $p > 0$, $\sigma_0 < \sigma < 1$ and $p < 0$, $-1 < \sigma < -\sigma_0$,

there are no solutions.

These formulae are valid for any p which is not itself so small as to be of order $1/r_0^2$. If they are applied to modes varying as $e^{jn\theta}$, where $n = \pm 1$ and σ , p are positive, they indicate the following results: for $\sigma > 1$, $n = \pm 1$, $\beta^2 \rightarrow 1 + \frac{p}{\sigma - 1}$; for $\sigma_0 < \sigma < 1$, no $n = \pm 1$ modes; for $0 < \sigma < \sigma_0$, $n = \pm 1$, $\beta^2 \rightarrow 1 + \frac{p}{\sigma + 1}$. These, in turn, may be classified in the following way. For $\sigma > 1$, $n = -1$, and for $0 < \sigma < \sigma_0$, $n = +1$ which correspond to μ and κ both positive, the propagation constant tends to the value for a plane wave whose direction of circular polarization coincides with that of the wave guide pattern. For $\sigma_0 < \sigma < 1$,

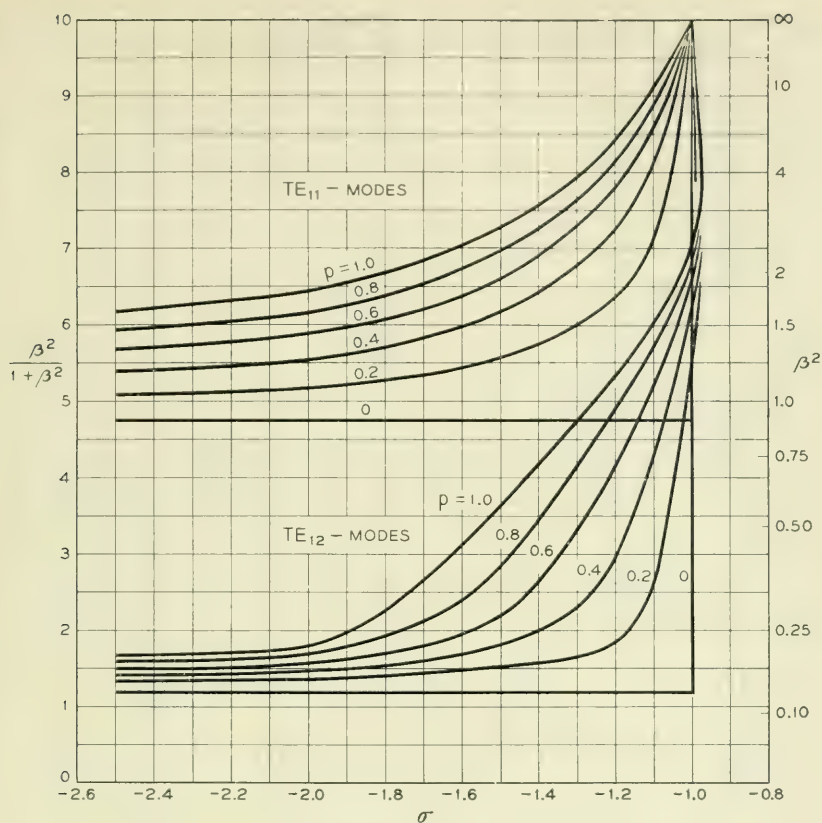


Fig. 13(f) — See Fig. 13.

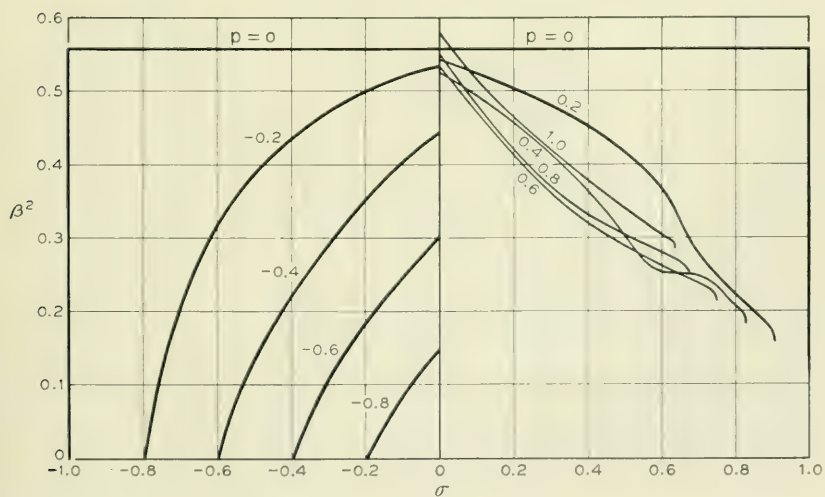


Fig. 13(g) — See Fig. 13.

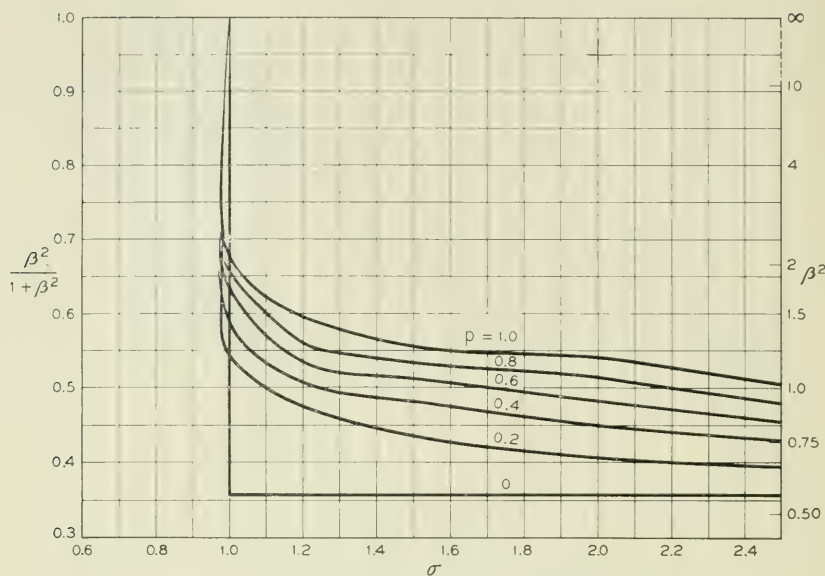


Fig. 13(h) — See Fig. 13.

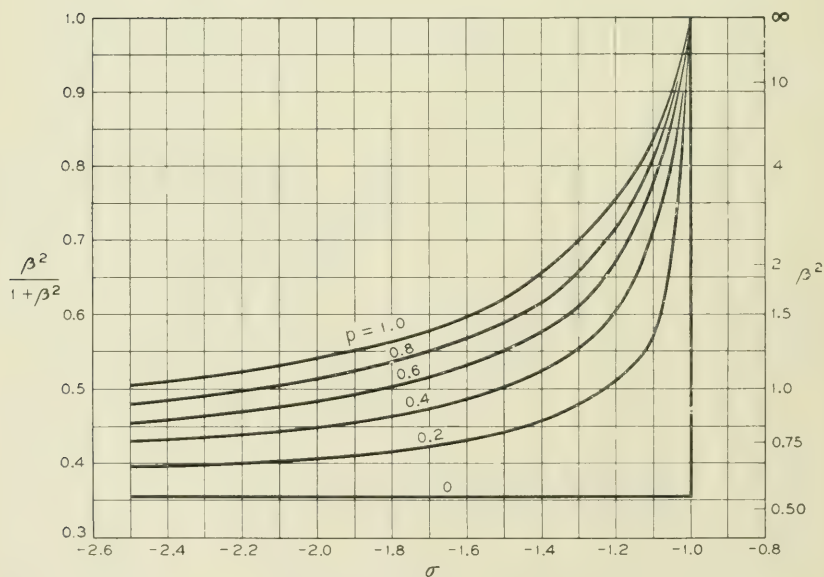


Fig. 13(i) — See Fig. 13.

where μ is negative, $n = \pm 1$, no modes exist for large enough guide. For $0 < \sigma < \sigma_0$, $n = -1$ and $\sigma > 1$, $n = +1$, the propagation constant tends to that for a plane wave whose polarization is in the opposite sense to that of the field pattern and here μ is positive, but κ is negative. An examination of the field pattern in this last case shows that most of the field energy is indeed associated with a circular polarization opposite to that of the pattern as a whole.

The discussion of the preceding sections shows that the complete structure of the mode spectrum for a guide filled with lossless ferrite is very complex. It is also clear that for some combinations of guide size and magnetic parameters the course of an individual mode in the $\beta^2 - \sigma$ plane may be quite involved. In particular, two values of β^2 associated with the same mode often occur at a given σ . The extent to which the complexity of the spectrum will be observed in practice will depend principally upon the loss of the real ferrite and upon the guide radius. The effect of loss near $\sigma = 1$, where the incipient modes are crowded will be to cause simultaneous excitation of many of these and consequently a confused z dependence of the guide excitation. For values of v_0 just below j_n , the point of escape of the TE modes, the latter exist over considerable ranges of σ , see Fig. 12(a), and would probably be observable. The TE modes near u_n also persist over a wide range, but are double-valued. Concerning such double-valued waves it may be observed that from the results of the subsequent treatment of losses, it is clear that if $\frac{d\beta^2}{d|\sigma|} > 0$, it is necessary to put the source of power at the opposite end of the guide.

4.15. *Losses, Faraday rotation and merit figure.* So far the analysis has been concerned with the loss-free medium. It is of some interest to determine the attenuation constant (the imaginary part of β) that arises when losses are taken into account. As long as these are small, this can be done rather easily; in fact, sufficiently far from resonance ($\sigma = 1$), for each formula giving β^2 , we can establish one giving the attenuation constant.

If the losses are of magnetic origin we utilize the fact (already demonstrated in section 2) that to first order in α , the permeabilities μ , κ are functions of $\sigma + j\alpha \operatorname{sgn} p$, and of no other combination of σ , α . Since σ , α enter Maxwell's equations only through μ and κ , β^2 , which is derived from them, must likewise depend on σ through $\sigma + j\alpha \operatorname{sgn} p$. Any formula for β^2 derived for the loss-free medium can, therefore, be generalized to the lossy case by replacing σ with $\sigma + j\alpha \operatorname{sgn} p$, to first order in α . To this order, then, we find

$$\beta^2 = \beta'^2 + ja \operatorname{sgn} p \frac{\partial(\beta')^2}{\partial \sigma},$$

where β' is the propagation constant for the loss-free case. Thus

$$j\beta = j\beta' - \frac{\alpha}{2\beta'} \frac{\partial(\beta')^2}{\partial |\sigma|}, \quad (44)$$

and the last term on the right, (multiplied by our scaling variable $\beta_0 = \omega\sqrt{\mu_z\epsilon_0}$) is the attenuation in nepers per meter. The present convention is that the waves propagate in the positive z direction, as $\exp(-j\beta z)$. It follows that they will decrease in that direction only if $\partial(\beta')^2/\partial|\sigma| < 0$. Occasionally this is not the case, and presumably indicates that the direction of the power flow opposes that of the phase velocity.

For small dielectric loss, too, it is possible to derive formulae for the attenuation constant from those already obtained; obviously the latter depend on ϵ only through $\epsilon = \epsilon_0 - j\epsilon_1$, and can therefore be expanded. But it must now be remembered that β was defined as $\beta_{\text{actual}}/\omega\sqrt{\mu_z\epsilon}$, and r_0 as $r_{\text{actual}}\omega\sqrt{\mu_z\epsilon}$ so that the scaling parameter $\omega\sqrt{\mu_z\epsilon}$ will make contributions to the imaginary part. It is then readily verified that

$$\frac{\beta_{\text{actual}}}{\omega\sqrt{\mu_z\epsilon_0}} = \beta' - j \frac{\epsilon_1}{2\epsilon_0} \frac{1}{\beta'} \frac{\partial}{\partial(r_0^2\beta^2)}. \quad (45)$$

A few words may be said about the relation of Faraday rotation to the $\beta^2 - \sigma$ curves. A linearly polarized plane wave traveling in the unbounded medium along the magnetizing field can be regarded as the

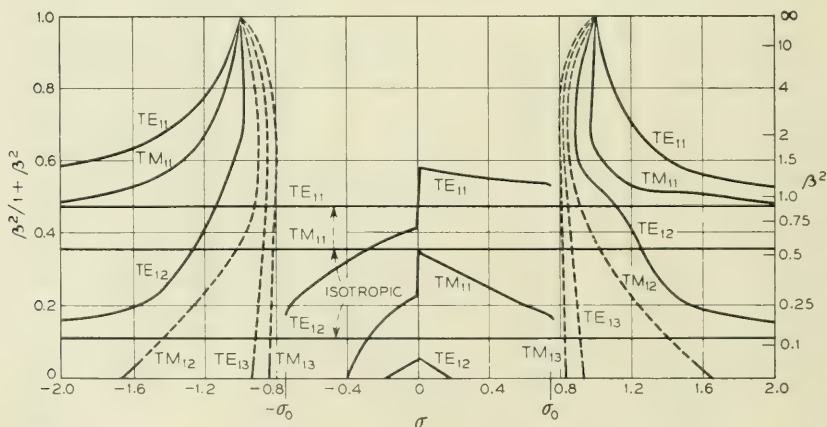


Fig. 14 — The course of the fully developed modes (solid lines) and of some of the lower incipient modes (dotted lines) as a function of σ for $r_0 = 5.75$ and $|p| = 0.6$.

sum of right and left circular components which travel with different propagation constants. If these are β_+ and β_- (measured in units of β_0) the plane of polarization of the resultant will appear to rotate by $(\beta_+ - \beta_-)/2$ radians per reduced wavelength $\frac{\lambda}{2\pi} = \frac{1}{\beta_0}$.

In the filled waveguide, on the other hand, it is no longer true that right and left circularly polarized modes add up to a plane polarized mode, as is readily seen by reference to the field components given in Appendix IV. To define Faraday rotation in a simple way it is therefore necessary to neglect changes in the field pattern due to the magnetization and consider only the changes in the propagation constants. Then the rotation of a mode with azimuthal mode number n will be $\frac{1}{2n}(\beta_+ - \beta_-)$.

In the present case $n = 1$, and the β_+ , β_- are found from the curves of β^2 versus σ , Figs. 13(a) to (i), for positive (σ, p) and negative (σ, p) respectively.

The merit figure is defined as the ratio — radians rotation per neper loss — and is independent of path-length. For small losses, (neglecting terms $O(\alpha^2)$), this ratio is

$$\frac{Rl(\beta_+ - \beta_-)}{Im(\beta_+ + \beta_-)} = \frac{1}{\alpha} \frac{\beta_+' - \beta_-'}{\frac{\partial \beta_+'}{\partial |\sigma|} + \frac{\partial \beta_-'}{\partial |\sigma|}},$$

in the notation of the present Section.

4.16. Formulae for the ferrite.

I. Cut-off points

Cut-off points will be classified into three types, 1, 2 and 3, according to the nature of the intersecting curves which generate them. All points of a given type may be assigned an index which further identifies the generating curve. This will be written as a subscript.

Type 1. Intersections of $I_n - (I_A')_T$, $\sigma > 0$, written as 1_n and of $I_n' - (I_A)_T$, $\sigma < 0$, written as $1_n'$
 $\beta^2 = 0$

There is a parametric representation:

$$\begin{aligned} |\lambda_{1,2}| &= e^\theta, \\ |p| &= 2 \sinh \theta \left(1 - \frac{r_0^2}{j_n^2} \right) \text{ and} \\ |\sigma| &= \frac{r_0^2}{j_n^2} e^\theta + \left(1 - \frac{r_0^2}{j_n^2} \right) e^{-\theta}. \end{aligned} \quad (46)$$

The slope at cut-off:

$$\left(\frac{\partial \beta^2}{\partial \sigma}\right)_p = \frac{\frac{j_n^2}{r_0^2} \left(\frac{j_n^2}{r_0^2} - 1\right) \coth \theta}{\left[\frac{j_n^2}{r_0^2} - 1 + \frac{2}{F(r_0)}\right] e^{-\theta} - e^{\theta}} \cdot \operatorname{sgn} \sigma, \quad (47)$$

and

$$\left(\frac{\partial \beta^2}{\partial p}\right)_\sigma = -\frac{\sigma}{p \coth \theta} \left(\frac{\partial \beta^2}{\partial \sigma}\right)_p. \quad (48)$$

Type 2. Intersections of

$$(0_n')_T - 0_c \text{ at } \sigma = \sigma_0 = \sqrt{\frac{p^2}{4} + 1} - \frac{p}{2}; \text{ type } 2_n'$$

$$(0_n)_T - 0_c' \text{ at } \sigma = -\sigma_0; \text{ type } 2_n.$$

$$\beta^2 \neq 0$$

Define λ_0 as a root of

$$\sigma_0 = \frac{-1}{|\lambda_0|} \left[1 - (1 - \lambda_0^2) \left(\frac{r_0}{F^{-1}(\lambda_0)} \right)^2 \right], \quad (49)$$

with the following convention: for $2_n'$, λ_0 is negative and the n^{th} branch of $F^{-1}(\lambda_0)$ is used; for 2_n , λ_0 is positive and the n^{th} branch of $F^{-1}(\lambda_0)$ is used. Now

$$\beta^2 = \frac{|\lambda_0|}{\sigma_0}. \quad (50)$$

Near cut-off

$$\beta^2 = \frac{|\lambda_0|}{\sigma_0} + \frac{r_0}{A} \sqrt{\sigma_0 \frac{1 + \sigma_0^2}{1 - \sigma_0^2}} \sqrt{\frac{\sigma_0 - |\sigma|}{1 + |\lambda_0| \sigma_0}}. \quad (51)$$

with

$$(1 - \lambda_0^2)A = \frac{1}{2|\lambda_0|} \left(1 - \frac{r_0^2}{1 + |\lambda_0| \sigma_0} \right) (\sigma_0 + 2|\lambda_0| + \sigma_0 \lambda_0^2) + \frac{1 + \sigma_0 |\lambda_0|}{\lambda_0}.$$

For 2_0 a special situation arises for $1 < r_0 < u_1$ where there are two positive solutions for λ_0 . We write 2_{01} and 2_{02} for the points corresponding to the smaller and greater of these. A special cut-off point will be labeled $2_{0\infty}$ and arises from $0_c'$ and $(0_0)_T$ as λ_1 goes to infinity. For this point

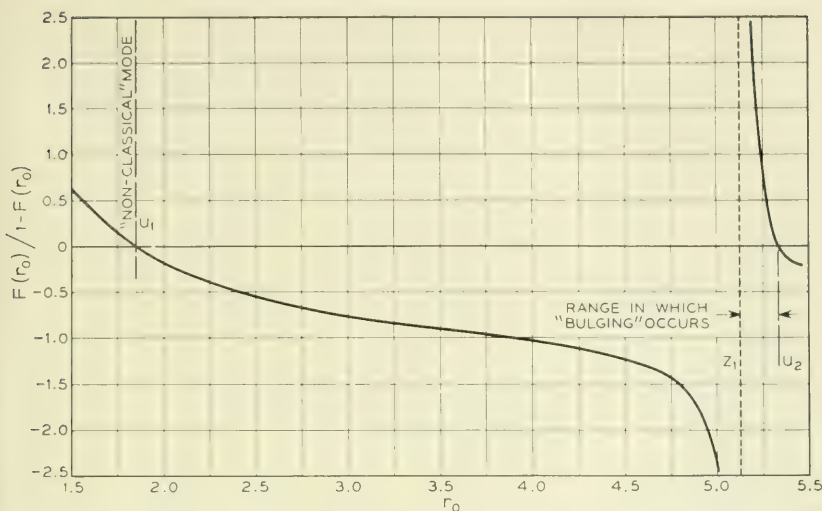


Fig. 15 — The function $\frac{F(r_0)}{1 - F(r_0)}$, related to cutoff of Type 3.

$\tau = \sigma_0$, $\beta^2 \rightarrow \infty$ and near $-\sigma_0$ we have

$$\beta^2 = \frac{|p|}{r_0^2 (|\sigma_0| - |\sigma|)} \frac{1}{(2 - |p| |\sigma_0|)}. \quad (52)$$

Type 3. Intersections of $I_B - (I_A)_T$; for $-\sigma_0 < \sigma < 0$ only; no subscript is needed.

$$\beta^2 = 0,$$

$$\sigma = -1 - p,$$

and

$$\frac{(\partial \beta^2)}{(\partial \sigma)_p} = \frac{(\partial \beta^2)}{(\partial p)_\sigma} = \frac{1}{p} \frac{F(r_0)}{1 - F(r_0)} \quad (\text{see Fig. 15}) \quad (54)$$

The cut-off points of the modes follow various schemes in different ranges of σ as indicated below.

For $\sigma > \sigma_0$ we have Table I. When $\sigma < -\sigma_0$, $1_n'$ replaces 1_n in the Table I. "None" indicates that the mode exists, but has no cut-off.

For $0 < \sigma < \sigma_0$ we have Table II. "N.P." in Table II indicates that the mode is not propagated. For $-\sigma_0 < \sigma < 0$ we have Table III. In this range of σ one has also the mode without classical analogue. For $r_0 < u_1$ this is cut-off at 3 and 2_{01} and may have a second branch from 2_{02} to $2_{0\infty}$. For $r_0 > u_1$ the second branch only exists.

TABLE I

Mode Radius	TE ₁₁	TM ₁₁	TE ₁₂	TM ₁₂	TE ₁₃ etc.
$r_0 < u_1$	1 ₁	1 ₂	1 ₃	1 ₄	1 ₅
$u_1 < r_0 < j_1$	None	1 ₁	1 ₂	1 ₃	1 ₄
$j_1 < r_0 < u_2$	None	None	1 ₂	1 ₃	1 ₄
$u_2 < r_0 < j_2$	None	None	None	1 ₂	1 ₃
$j_2 < r_0 < u_3$ etc.	None	None	None	None	1 ₃

II. Asymptotes, genesis of the modes and spot points

For $|\sigma| \rightarrow \infty$ there are asymptotic formulae:

For TE_{1n}-modes

$$\beta^2 \rightarrow \left(1 - \frac{u_n^2}{r_0^2}\right) \left(1 + \frac{p}{\sigma}\right). \quad (55)$$

For TM_{1n}-modes

$$\beta^2 \rightarrow \left(1 - \frac{j_n^2}{r_0^2}\right) + \frac{p}{\sigma}. \quad (56)$$

For $|\sigma| > \sigma_0$, all modes have their origin in the points $\sigma = 1, \lambda_1 = 1$ or $\sigma = -1, \lambda_2 = -1$, where $\beta^2 \rightarrow \infty$. The variation of β^2 with σ in the neighborhood of these points is described by the two expansions:

for $\sigma \sim 1$

$$\beta_n^2 = \frac{p}{a_n + 1} \left[\frac{1}{\sigma - 1} + \left\{ \frac{1 - a_n}{p} + \frac{2a_n^2}{x_n^2} \left(\frac{2}{p} + 1 \right) + \frac{a_n^2}{2} \right\} + 0(\sigma - 1) \right], \quad (57)$$

where

$$\frac{a_n}{a_n + 1} = \frac{x_n^2}{2r_0^2},$$

and

$$F(x_n) = 1 \quad x_n > 0,$$

with the scheme

n	1	2	3	4 etc.
Mode	TE ₁₁	TM ₁₁	TE ₁₂	TM ₁₂ etc.

TABLE II

Mode Radius	TE ₁₁	TM ₁₁	TE ₁₂	TM ₁₂	TE ₁₃ etc.
0 < r ₀ < u ₁	N.P.	N.P.	N.P.	N.P.	N.P.
u ₁ < r ₀ < j ₁	2 ₀ '	N.P.	N.P.	N.P.	N.P.
j ₁ < r ₀ < u ₂	2 ₀ '	1 ₁	N.P.	N.P.	N.P.
u ₂ < r ₀ < j ₂	2 ₀ '	2 ₁ '	1 ₁	N.P.	N.P.
j ₂ < r ₀ < u ₃	2 ₀ '	2 ₁ '	1 ₁	1 ₂	N.P.
u ₃ < r ₀ < j ₃ etc.	2 ₀ '	2 ₁ '	2 ₂ '	1 ₁	1 ₂

TABLE III

Mode Radius	TE ₁₁	TM ₁₁	TE ₁₂	TM ₁₂	TE ₁₃ etc.
0 < r ₀ < u ₁	N.P.	N.P.	N.P.	N.P.	N.P.
u ₁ < r ₀ < j ₁	3	N.P.	N.P.	N.P.	N.P.
j ₁ < r ₀ < u ₂	3	1 ₁ '	N.P.	N.P.	N.P.
u ₂ < r ₀ < j ₂	2 ₁	3	1 ₁ '	N.P.	N.P.
j ₂ < r ₀ < u ₃	2 ₁	3	1 ₁ '	1 ₂ '	N.P.
u ₃ < r ₀ < j ₃ etc.	2 ₁	2 ₂	3	1 ₁ '	1 ₂ '

for $\sigma \sim -1$

$$\beta_n^2 = \frac{p}{a_n + 1} \left[\frac{1}{\sigma + 1} + \left\{ \frac{1 - a_n}{p} + \frac{2a_n^2}{y_n^2} \left(1 - \frac{2}{p} \right) - \frac{a_n^2}{2} \right\} + 0(\sigma + 1) \right], \quad (58)$$

where

$$\frac{a_n}{a_n + 1} = \frac{y_n^2}{2r_0^2},$$

and

$$F(y_n) = -1,$$

with the same identification as above.

For $\sigma > \sigma_0$ a spot-point is given by $I_n - (I_B)_T$ with

$$\beta^2 = \lambda_1 = 1 + \frac{p}{1 + \sigma},$$

and

$$\frac{1 - \lambda_1^2}{1 - \sigma \lambda_1} = \frac{j_n^2}{r_0^2},$$

The identification scheme is again that shown above.

For $-\sigma_0 < \sigma < 0$ an isolated identifiable point arises from I_B' - $(I_n)_T$ which is expressible in inverse form by the relations

$$\sigma = \frac{1 - \frac{r_0^2}{j_n^2}}{\beta^2} + \frac{r_0^2}{j_n^2} \beta^2 \quad \text{and} \quad p = (\beta^2 - 1) \left[1 + \frac{r_0^2}{j_n^2} \beta^2 + \frac{1 - \frac{r_0^2}{j_n^2}}{\beta^2} \right] \quad (60)$$

for

$$1 - \frac{j_n^2}{r_0^2} < \beta^2 < \sqrt{1 - \frac{j_n^2}{r_0^2}}.$$

The identification of the modes with n proceeds as in the earlier parts of this section.

III. Small p .

To order p^2 there exist the following expansions for β^2 : for the TE_{1n} -mode

$$\beta^2 = \beta_0^2 + \frac{\beta_0^2}{1 - \sigma^2} \left[\frac{2}{u_n^2 - 1} - \sigma \right] p + \frac{2\beta_0^2}{(1 - \sigma^2)^2} \cdot \frac{\beta_0^2}{(u_n^2 - 1)^2} + \frac{5 + 5u_n^2 - 2u_n^4}{4(u_n^2 - 1)} \frac{\beta_0^2}{(\beta_0^2 - 1)(1 - u_n^2)} p^2 \dots, \quad (61)$$

where

$$\beta_0^2 = 1 - \frac{u_n^2}{r_0^2}.$$

For the TM_{1n} -modes

$$\beta^2 = \beta_0^2 - \frac{\sigma}{1 - \sigma^2} p + \frac{1}{(1 - \sigma^2)^2} \frac{3\beta_0^2 - 2}{2(1 - \beta_0^2)} p^2 \dots, \quad (62)$$

where

$$\beta_0^2 = 1 - \frac{j_n^2}{r_0^2}.$$

The radius of convergence of these series is not known. It is clear that it will depend on σ and will become smaller as $\sigma^2 \rightarrow 1$.

4.2. *The Plasma* ($\rho_H = 0$, $\nu_H = 1$). The characteristic equation (26) may now be written

$$\frac{1}{\chi_{1,2}^2} [\lambda_1 F_n(\chi_1 r_0) - n] = \frac{1}{\chi_{1,2}^2} [\lambda_2 F_n(\chi_2 r_0) - n], \quad (63)$$

where (in contrast with the ferrite case) $\lambda_{1,2} = \beta \Lambda_{1,2}$. The λ satisfy

$$\lambda_{1,2}^2 - \frac{(\nu_E - 1)(1 - \beta^2/\nu_E) - \nu_E \rho_E^2}{\rho_E} \lambda_{1,2} - \beta^2 = 0, \quad (64)$$

and the χ 's are given by

$$\chi_{1,2}^2 = (1 - \beta^2/\nu_H) - \rho_H \lambda_{1,2} \quad (65)$$

From the equations for ρ_E and ν_E in terms of σ , q given in Section 2, equation (64) may be written

$$\lambda_{1,2}^2 - \left(\frac{\sigma}{1 - q^2} - \sigma \beta^2 \right) \lambda_{1,2} - \beta^2 = 0, \quad (66)$$

or

$$\lambda_1 \lambda_2 = -\beta^2 \quad (67a)$$

$$\begin{aligned} \lambda_1 + \lambda_2 &= \frac{\sigma}{1 - q^2} - \sigma \beta^2, \\ &= \frac{\sigma}{1 - q^2} + \sigma \lambda_1 \lambda_2. \end{aligned} \quad (67b)$$

Elimination of β^2 between equations (64) and (65) enables us to express $\chi_{1,2}^2$ solely in terms of $\lambda_{1,2}$, ρ_H and ν_H :

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \frac{1 - 1/\nu_E}{\rho_E} \lambda_{1,2}},$$

which, from the plasma formulae for ρ_E , ν_E , can be written

$$\chi_{1,2}^2 = \frac{1 - \lambda_{1,2}^2}{1 - \sigma \lambda_{1,2}}. \quad (68)$$

With these expressions for the χ , the characteristic equation (63) takes the form

$$H(\lambda_1, \sigma, r_0) = H(\lambda_2, \sigma, r_0),$$

where

$$H(\lambda, \sigma, r_0) = \frac{1 - \sigma\lambda}{1 - \lambda^2} \left[\lambda F_n \left(r_0 \sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}} \right) - n \right]. \quad (69)$$

For given σ , and q , equations (67b) and (69) are simultaneous equations for λ_1, λ_2 . When λ_1, λ_2 have been found, $\beta^2 = -\lambda_1\lambda_2$ is known. Since β^2 must be positive, λ_1, λ_2 must have opposite signs. As in the ferrite, the convention $\lambda_1 > 0, \lambda_2 < 0$ will be adopted. Equation (67b) will hereafter be called the plasma relation. The transformation

$$\lambda_1 \rightarrow -\lambda_2, \quad \lambda_2 \rightarrow -\lambda_1, \quad \sigma \rightarrow -\sigma$$

leaves the plasma relation unchanged and changes n to $-n$ in the H-equation. As more fully explained in connection with the ferrite section, it is therefore necessary to consider positive n only if σ is allowed to take on negative as well as positive values. As before, only the first azimuthal mode number ($n = \pm 1$) is considered in this paper.

The method of analysis is the same as that used for the ferrite. Here we shall only sketch the most important steps; the reader will have no difficulty in completing the analysis by referring to Section 4.11. For fixed r_0 , a contour map of H is drawn in the λ, σ plane (see Fig. 16 drawn for $r_0 \sim 2.2$). The gross features of this map are determined by the lines $H = 0, H = \pm \infty$. For greater detail recourse is had to the lines $\frac{1 - \lambda^2}{1 - \sigma\lambda} = \text{constant}$, along which values of H are readily generated. Further help is obtained from a knowledge of the location of the saddle point of H . The infinity curves are given by the same formulae as for the ferrite, except that the line $\lambda = 0$ is no longer an infinity line: along $\lambda = 0, H = -1$. Zero curves are given by

$$\sigma = \frac{1}{\lambda} - \frac{r_0^2(1 - \lambda^2)}{\lambda[F^{-1}(\lambda^{-1})]^2}.$$

The branches of $\sigma\lambda = 1$ are also zero curves in the same restricted sense as for the G function. In the same notation as for the ferrite, all I_n curves pass through $\sigma = 1, \lambda = 1$; all I_n' curves through $-1, -1$. The same is true for all $0_n, 0_n'$ curves ($n > 0$). The only exception is denoted by 0_0 , it arises from that branch of F^{-1} along which $F^{-1}(1) = 0$.

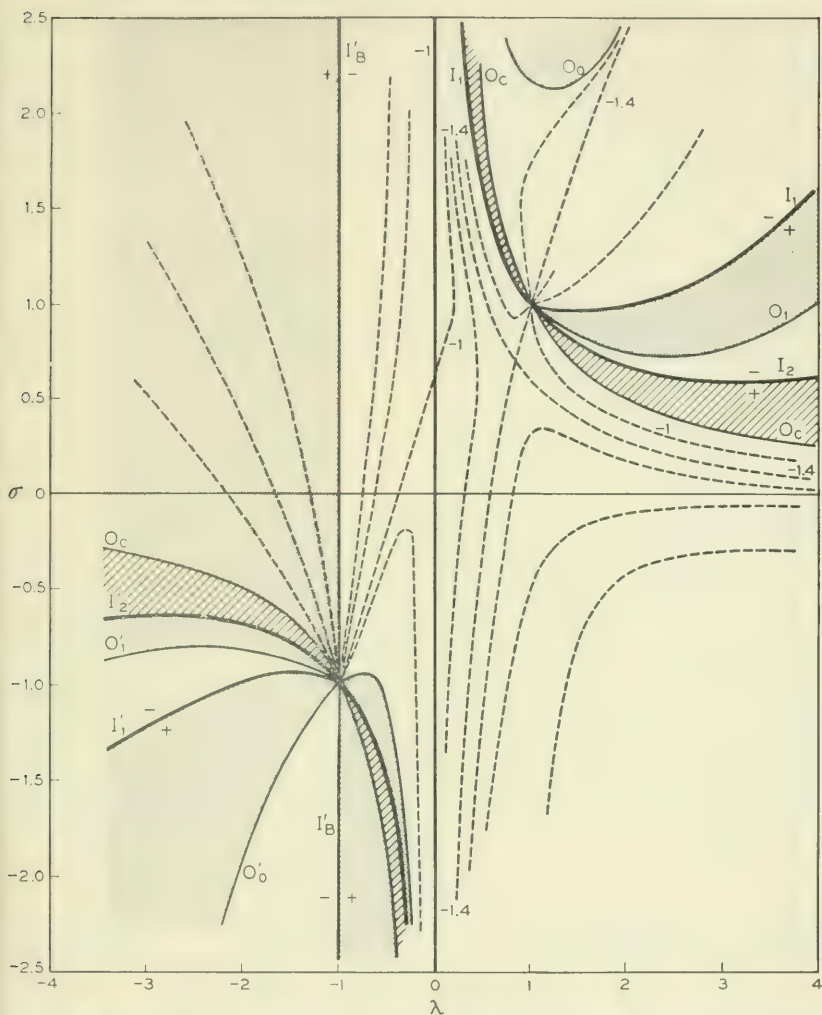


Fig. 16 — The division of the $\lambda - \sigma$ plane into regions of positive and negative H by the first few O and I curves. Dotted regions are positive. Cross-hatched regions contain the higher O and I curves.

Along all lines, $\sigma = \lambda c + d$, H tends to infinity except when $c = r_0^2/u_n^2$ (the slopes of the linear asymptotes of O curves). Along $\sigma = \frac{r_0^2}{u_n^2} \lambda + d$, H tends to a value depending on d . More about the general behavior of H can be derived by means entirely analogous to those employed in the study of G .

λ -pairs determined from the H -diagram do not necessarily solve the problem, since for a fixed q , they may not satisfy the plasma relation (67b). To take it into account, we interpret it as a transformation of the whole of the second quadrant onto part of the first, and of the whole fourth quadrant on part of the third. Writing (67b) in the form

$$\lambda_1 = \frac{1}{\sigma} + \frac{\frac{\sigma^2}{1 - q^2} - 1}{\sigma \frac{1 - \sigma \lambda_2}{1 - q^2}} = T(\lambda_2),$$

we see that the curves $\lambda_2 = \text{const.}$ transform into a bundle of hyperbola passing through the intersection of $\sigma^2 = 1 - q^2$ with $\sigma = 1/\lambda$, that is, through

$$\lambda_{10} = \frac{1}{\sqrt{1 - q^2}}, \quad \sigma_0 = \sqrt{1 - q^2}.$$

These hyperbolae have vertical asymptotes $\lambda_1 = -\frac{1}{\lambda_2}$, and cut the line $\sigma = 0$ in $-\lambda_2$. For a fixed positive $\sigma < \sigma_0$, λ_1 decreases from $1/\sigma$ to $\sigma/(1 - q^2)$ as λ_2 increases from $-\infty$ to 0, but, when $\sigma > \sigma_0$, λ_1 increases from $1/\sigma$ to $\sigma/(1 - q^2)$. Thus the second quadrant transforms into the region between $\sigma = \lambda(1 - q^2)$ and $\sigma = 1/\lambda$ in the first quadrant. Similarly, the inverse transformation $\lambda_2 = T(\lambda_1)$ transforms the fourth quadrant into the region between $\sigma = \lambda(1 - q^2)$ and $\sigma = 1/\lambda$ in the third quadrant. Points outside these regions cannot be site of acceptable solutions of the H equation. In order to locate acceptable solutions, the $H = \text{equation}$ is now written in the form

$$H(\lambda_1, \sigma, r_0) = H(T(\lambda_1), \sigma, r_0)$$

when $\sigma > 0$, and in the form

$$H(\lambda_2, \sigma, r_0) = H(T(\lambda_2), \sigma, r_0)$$

when $\sigma < 0$. These equations represent the curves of intersection of the H -surfaces. Along each such curve, both H -equation and plasma relation are satisfied. Their projections onto the first (or third) quadrant give λ_1 (or λ_2) as a function of σ , and hence λ_2 (or λ_1) from the plasma relation. Thus $\beta^2 = -\lambda_1\lambda_2$ is known along each solution curve. The rough locating of the solution curves, and the establishment of precise analytical formulae near special points on them proceeds in complete analogy with the ferrite case. Here we shall consider only the radius $r_0 \sim 2.2$, as typical of radii large enough to permit propagation of the TE_{11} mode

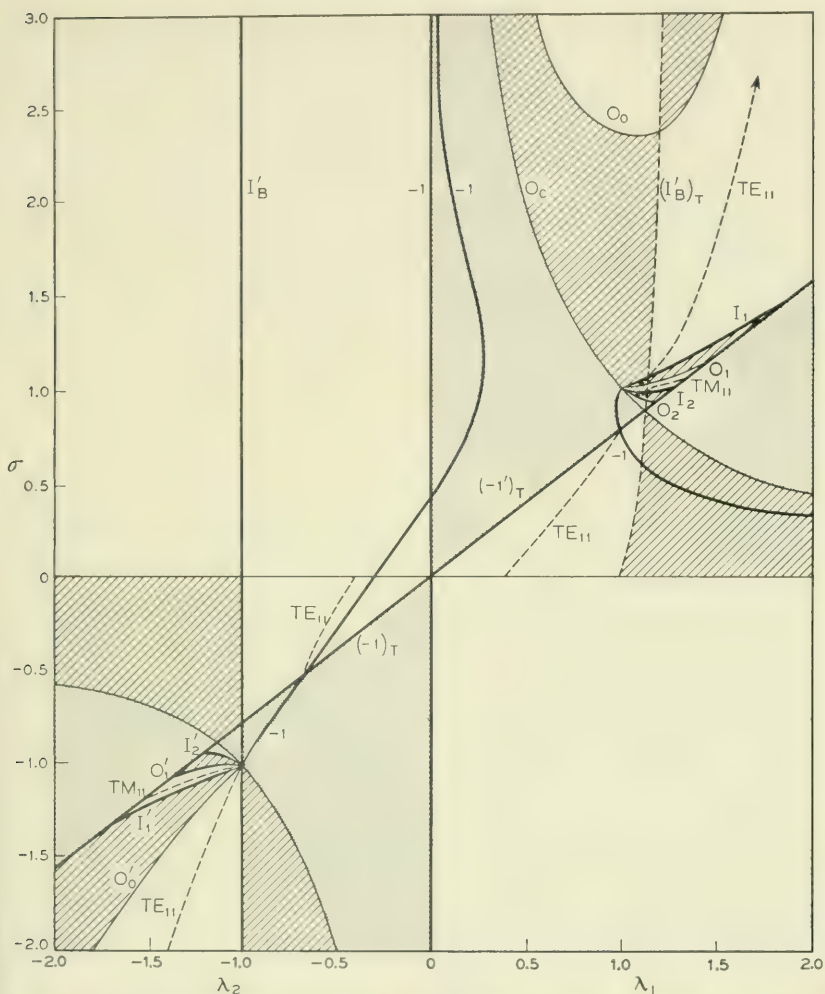


Fig. 17 — Geometrical exploration of solution curves for the plasma. The dotted regions are excluded by the plasma relation; the cross-hatched regions are those in which $H(\lambda, \sigma)$ and $H(T(\lambda), \sigma)$ have unlike sign. Solution curves may lie only in unshaded parts of the 1st and 3rd quadrants. $q^2 = 0.25$, $r_0 \sim 2.0$.

through the unmagnetized plasma, but too small to admit higher modes. The solution curves for $r_0 \sim 2.2$ are indicated roughly in Fig. 17.

In the first quadrant, for $\sigma > \sigma_0$, a solution curve starts at $\lambda_1 = 1$, $\sigma = 1$, passes through the intersection of I_1 , $(I'_B)_T$ and proceeds to infinity as indicated. The formulae in Section 4.21 describe the corresponding curve near $\sigma = 1$, and at $\sigma \rightarrow \infty$, showing that the solution

curve describes the TE_{11} -limit mode. Incipient modes also exist, just as in the ferrite; their end-points on $\sigma = (1 - q^2)\lambda_1$ (or, briefly, on $(\lambda_2 = 0)_T$) are now the points for which $H(\lambda_1, \sigma) = -1$ and, simultaneously, $\sigma = (1 - q^2)\lambda_1$.

Below σ_0 , there is only one solution curve for $r_0 \sim 2.2$. It begins at $\sigma = 0$, $\lambda_1 = \beta_{iso}$ ($= -\lambda_2$, by the plasma relation), where β_{iso} is the propagation constant of the TE_{11} mode in the unmagnetized plasma. (In contrast with the ferrite, the plasma becomes isotropic as $\sigma \rightarrow 0$). It is cut off at the intersection of the contour $H(\lambda_1, \sigma) = -1$ in that region with $(\lambda_2 = 0)_T$. At that point $\beta^2 = 0$ and σ is best stated, thus:

$$\sigma = (1 - q^2) \sqrt{(1 + y^2)/[1 + (1 - q^2)y^2]},$$

where y is the (unique) real root of

$$F(jyr_0) = \sqrt{(1 + y^2)[1 + (1 - q^2)y^2]}.$$

Alternatively these two equations may (by varying y) be used to generate r_0 's and the corresponding cut-off values of σ . Of course, the two equations are merely a re-statement of the equations $H(\lambda_1, \sigma) = -1$, $\sigma = \lambda_1(1 - q^2)$, heed being paid to the fact that the argument of F is imaginary in the region considered for the radius under discussion.

In the third quadrant for $\sigma < -\sigma_0$, we also find the TE_{11} -limit mode. Its solution curve begins at $\lambda_2 = -1$, $\sigma = -1$, and proceeds to $\sigma = -\infty$ without passing through any easily computed intersections of I curves. Formulae pertaining to the TE_{11} mode in this range are stated in Section 4.22. Again the incipient modes are found in their usual region. For $0 > \sigma > -\sigma_0$, the solution curve corresponding to the TE_{11} mode begins at $\sigma = 0$, $\lambda_2 = -\beta_{iso}$ ($= -\lambda_1$) and is cut off at the intersection of $H(\lambda_2, \sigma) = -1$ with $\lambda_2(1 - q^2) = \sigma$ (or $(\lambda_1 = 0)_T$). At that point $\beta^2 = 0$, and σ is given by

$$\sigma = -(1 - q^2) \sqrt{(1 - y^2)/[1 - (1 - q^2)y^2]},$$

where y is the least real root of

$$F(r_0y) = -\sqrt{(1 - y^2)[1 - (1 - q^2)y^2]}.$$

Alternatively, this equation can be used to generate r_0 , and the associated σ , if y is regarded as a parameter, which for $u_1 < r_0 < j_1$ is between zero and unity.

At a fixed r_0 the higher roots of the last equation with sign reversed and the corresponding σ are associated with the cut-offs of the incipient

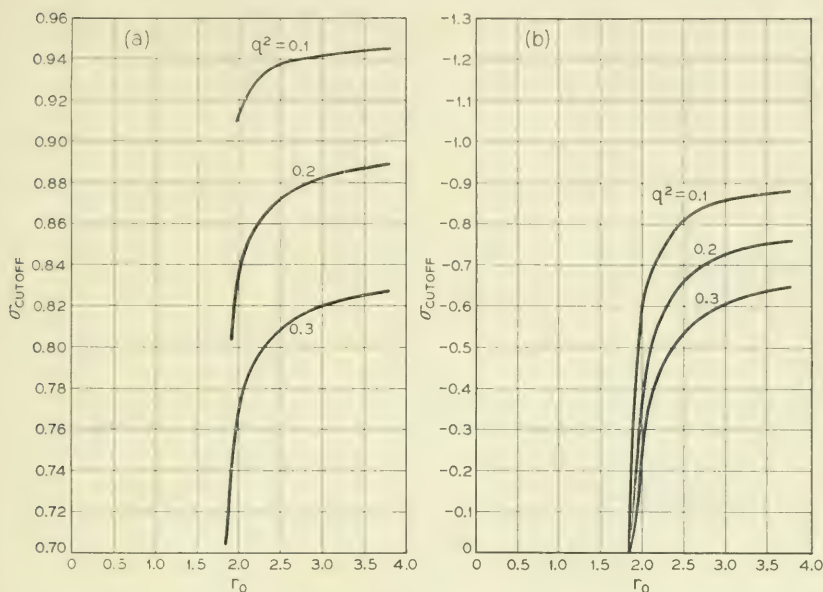


Fig. 18 — Cutoff value of σ for the TE_{11} -limit mode in the plasma as a function of r_0 for various q^2 .

modes in the lower half plane. Similarly the real roots of

$$F(r_0 y) = -\sqrt{(1 - y^2)[1 - (1 - q^2)y^2]},$$

and the corresponding

$$\sigma = +(1 - q^2) \sqrt{(1 - y^2)/[1 - (1 - q^2)y^2]},$$

are associated with the cut-offs of the incipient modes in the upper half-plane. These equations have been solved for the TE_{11} mode and their solutions shown in Figs. 18(a) and 18(b).

4.21. Some formulae relating to the plasma (chiefly for TE-modes).

The formulas given here employ dimensionless variables (Section 3) except where otherwise stated.

Approximations for extreme values of σ or q^2

(a) small σ , q^2 not near unity

TE_{1m} mode:

$$\beta^2 = \beta_m^2 + A_m \sigma + B_m \sigma^2 + \dots \quad (70)$$

where

$$\beta_m^2 = 1 - \frac{u_m^2}{r_0^2},$$

$$A_m = \frac{2}{u_m^2 - 1} \frac{q^2}{1 - q^2}, \quad \text{and}$$

$$B_m = -\frac{q^2}{(1 - q^2)^2} \left[1 + \frac{q^2}{u_m^2 - 1} \left\{ \left(\frac{1}{2u_m^2} - 1 \right) \beta_m^2 + \frac{4}{(u_m^2 - 1)^2} \left(1 - \frac{1}{2u_m^2} \right) \right\} \right].$$

TM modes show no first order variation with σ .

(b) small q^2 , σ^2 not near unity. Here β_a and r_a are the actual propagation constant and radius, without scaling factors:

TE_{1m} mode:

$$\beta_a^2 = (1 - q^2) \omega^2 \mu_0 \epsilon_0 - \frac{u_m^2}{r_a^2} + \omega^2 \epsilon_0 \mu_0 \frac{\sigma q^2}{1 - \sigma^2} \left(\frac{2}{u_m^2 - 1} - \sigma \right) + O(q^4). \quad (71)$$

TM_{1m} modes:

$$\beta_a^2 = (1 - q^2) \omega^2 \mu_0 \epsilon_0 - \frac{j_m^2}{r_0^2} - \omega^2 \epsilon_0 \mu_0 \frac{\sigma^2 q^2}{1 - \sigma^2} \left(1 - \frac{j_m^2}{\omega^2 \mu_0 \epsilon_0 r_a^2} \right). \quad (72)$$

(c) Approximation for large σ ; q^2 not near unity.

TE_{1m} mode:

$$\beta^2 = \frac{1}{1 - q^2} - \frac{u_m^2}{r_0^2} - \frac{2q^2}{\sigma(1 - q^2)} \frac{1}{(u_m^2 - 1)}. \quad (73)$$

All formulae in (a), (b), (c) apply to both positive and negative σ (right and left circular waves). Formulae 70 and 73 show that the first order changes in β^2 , whether due to very large or very small σ , have coefficients that differ only in sign.

The TE₁₁ mode near resonance ($q^2 < 1$)

Near $\sigma = +1$,

$$\beta^2 = -\frac{q^2}{1 - q^2} \frac{\left(\frac{j_1^2}{2r_0^2} - 1 \right)}{\sigma - 1} \quad (74)$$

Near $\sigma = -1$,

$$\beta^2 = \frac{-q^2}{1 - q^2} \frac{1}{1 + \sigma}. \quad (75)$$

Cut-off of the TE_{11} mode ($q^2 < 1$) ($u_1 < r_0 < j_1$)

σ positive:

$$\beta^2 = 0; \quad \sigma = (1 - q^2) \sqrt{(1 + y^2/[1 + (1 - q^2)y^2])}, \quad (76)$$

where y is the only real root or the smallest imaginary root of

$$F(jr_0y) = \sqrt{(1 + y^2)[1 + (1 - q^2)y^2]}.$$

σ negative:

$$\beta^2 = 0; \quad \sigma = -(1 - q^2) \sqrt{(1 - y^2)/[1 - (1 - q^2)y^2]}, \quad (77)$$

where y is the smallest real root of

$$F(r_0y) = -\sqrt{(1 - y^2)[1 - (1 - q^2)y^2]}.$$

(See section 4.2 for further explanation).

APPENDIX I. THE F -FUNCTION

The function $F(x)$ has been defined by the equation

$$F(x) = x \frac{J_1'(x)}{J_1(x)}.$$

Using the infinite product for $J_1(x)$ and differentiating logarithmically one finds

$$F(x) = 1 - 2 \sum_{n=1}^{\infty} \frac{x^2}{j_n^2 - x^2}, \quad (78)$$

where $J_1(j_n) = 0$. Near one of its poles, j_n , $F(x)$ behaves as $j_n/(x - j_n)$. It is also useful to know the form of $F(x)$ near one of its zeros, u_n , which are also zeros of $J_1'(x)$. Such an expansion may conveniently be found by using the Ricatti equation satisfied by $F(x)$, which is

$$x \frac{dF}{dx} = 1 - x^2 - F^2. \quad (79)$$

The expansion near u_n is then

$$F(u_n + y) = y \left[\frac{1}{u_n} - u_n \right] - \frac{y^2}{2} \left[\frac{1}{u_n^2} + 1 \right] + \text{higher terms.} \quad (80)$$

TABLE IV

Equation Number	Asymptote	Polder Transform of Asymptote	Range of Validity
S3	$\sigma = -g \left[\lambda + F \left(\frac{r_0}{\sqrt{-g}} \right) \right]$	$\lambda = \frac{1+g}{\sigma} + O \left(\frac{1}{\sigma^2} \right)$	σ large, g finite and not equal to zero or $-r_0^2/j_n^2$
S4	$\sigma = \frac{r_0^2}{j_n^2} \left(\lambda - \frac{2}{1 + \frac{j_n^2}{r_0^2} g} \right) \quad n = 1, 2, 3, \dots$	Not required	σ large, g unrestricted
S5	$\sigma = \frac{1}{\lambda} + \frac{g^2}{r_0^2} \lambda$	Not required	Large σ Small λ $g > 0$ for $\lambda < 0$ $g < 0$ for $\lambda > 0$ g finite
S6	$\sigma = \left(1 - \frac{r_0^2}{u_n^2} \right) \frac{1}{\lambda} + \frac{2 \left(1 + \frac{u_n^2}{r_0^2} g \right)}{1 - u_n^2} \frac{r_0^2}{u_n^3}$ $n = 1, 2, 3, \dots$	$\sigma = \lambda \frac{r_0^2}{u_n^2} - p - \frac{2 \left(1 + \frac{u_n^2}{r_0^2} g \right)}{u_n(1 - u_n^2)} \left(1 - \frac{r_0^2}{u_n^2} \right)$	σ large g finite

The equation may also be used to furnish an expansion near $x = 0$. This is

$$F(x) = 1 - \frac{x^2}{4} - \frac{x^4}{96} + \text{higher terms.} \quad (81)$$

Finally, putting $x = jy$, one finds from Eq. (79) for large y

$$F(jy) = y - \frac{1}{2} + \frac{3}{8} \frac{1}{y} + \text{higher terms.} \quad (82)$$

APPENDIX II. INFORMATION PERTAINING TO THE CONSTRUCTION OF G -DIAGRAMS

The accurate construction of the contours $G = \text{const.}$ is conveniently based on the contours $(1 - \lambda^2)/(1 - \sigma\lambda) = \text{const.}$ along any one of which G is a function of λ alone. These contours are shown in Fig. 4. Their asymptotic properties are almost self-evident.

The curves $G = g = \text{const.}$ have various asymptotes. These, together with their range of validity, and their Polder transforms where needed are stated in Table IV.

The formulas given in Table IV show that the curves $G = \text{const.}$ generally have two kinds of asymptotes; linear and hyperbolic. Formula (83) shows the behavior of G along a line of constant finite slope unequal to r_0^2/j_n^2 , the asymptotic slope of the I_n curves. Parallel to a line of slope r_0^2/j_n^2 all G contours must be found, not just the restricted range given by the first formula. Writing $\sigma = (r_0^2/j_n^2)\lambda + x$ in the equation $G = g$, and expanding F near its pole j_n , we find x in terms of g and obtain (84) which holds for all g , from $-\infty$ to $+\infty$. When $g = 0$ it also gives the linear asymptotes of 0_n , $0_n'$ curves except 0_0 , as is readily verified from the equation

$$\sigma = \frac{1}{\lambda} - \frac{r_0^2(1 - \lambda^2)}{\lambda[F^{-1}(\lambda)]^2},$$

for the zero curves.

Formula (85) shows how the G -contours tend towards $\sigma\lambda = 1$ from the side $\sigma\lambda > 1$ as $\lambda \rightarrow 0$.

Formula (86) relates the asymptotic behavior of the curves $G = g$ to the zero curves, $g = 0$, for small λ . All G curves approach zero curves arbitrarily closely as $\lambda \rightarrow 0$. The only exceptions are the infinity curves whose form near $\lambda = 0$ is

$$\sigma = \frac{1}{\lambda} \left(1 - \frac{r_0^2}{j_n^2} \right) \quad (87)$$

When $r_0 = u_{n+1}$, the 0_n curve merges with the $0_n'$ curve at

$$\sigma = \frac{-2}{(u_{n+1}^2 - 1)} \frac{r_0^2}{u_{n+1}^3}.$$

Similarly all of the contours $G = g$ of the sheet to which 0_n belongs merge with the corresponding $G = g$ of the sheet to which $0_n'$ belongs, at

$$\sigma = -\frac{2(1+g)}{(u_{n+1}^2 - 1)} \frac{r_0^2}{u_{n+1}^3}. \quad (88)$$

These remarks apply to all $0_n, 0_n'$ curves, 0_0 included.

The 0_0 -curve, for large λ , behaves as

$$\sigma = \frac{1 - r_0^2}{\lambda},$$

and so tends to σ from above for $r_0 < 1$ and from below when $r_0 > 1$. (In fact, for $r_0 < 1$, 0_0 lies wholly in the first quadrant; when $u_1 > r_0 > 1$, 0_0 cuts $\sigma = 0$ once.)

The saddle points of G are most easily found by considering G in the coordinate net formed by the curves $\chi^2 = \text{const.}$ and $\lambda = \text{const.}$ At a saddle point

$$\frac{\partial G}{\partial \chi} = \frac{\partial}{\partial \chi} \left[\frac{1}{\chi^2} \left(\frac{1}{\lambda} F(r_0 \chi) - 1 \right) \right] = 0$$

and simultaneously

$$\frac{\partial G}{\partial \lambda} = 0$$

The only saddle points that might be missed in this way are points at which the two derivatives are not independent, that is points where the χ^2 contours have vertical tangents, and it is easily verified that no saddle points exist there.

Proceeding with the differentiations, we find that $\frac{\partial G}{\partial \lambda} = 0$ gives

$$F(r_0 \chi) = 0$$

or

$$r_0 \chi = u_n \quad (89)$$

and so $\frac{\partial G}{\partial \chi} = 0$ gives

$$\frac{2}{\chi} + \frac{r_0}{\lambda} F'(u_n) = 0$$

or

$$\lambda_{ns} = \frac{u_n^2 - 1}{2}. \quad (90)$$

The corresponding σ_{ns} are given by

$$\frac{1 - \lambda_{ns}^2}{1 - \sigma_{ns}\lambda_{ns}} = \frac{u_n^2}{r_0^2} \quad (91)$$

and are all positive. Thus all saddle points lie in the first quadrant. At a saddle point $G = -r_0^2/u_n^2$ and therefore it is the intersection of two contours $G = -r_0^2/u_n^2$. For $n > 1$, one of these obeys the asymptotic formula (83), the other is asymptotic to I_n and I_{n-1} (see Fig. 5), and obeys (84), with n and $n - 1$, near those curves. For $n = 1$, one of them still follows formula (83), but the two "arms" of the other are asymptotic to $\sigma = 1/\lambda$ and I_1 , and so follow (85) and (84) with $n = 1$ respectively.

Three further facts useful in the construction of G -diagrams are:

Along a curve $\frac{1 - \lambda^2}{1 - \sigma\lambda} = \frac{u_n^2}{r_0^2}$, G equals $-\frac{r_0^2}{u_n^2}$; thus the zero curves of F are contours of constant G .

Along $\lambda = +1$,

$$G = \frac{1 - \sigma}{2} - \frac{r_0^2}{4}. \quad (92)$$

As $\sigma, \lambda \rightarrow 1$ along $(\sigma - 1) = \alpha(\lambda - 1)$;

$$G \rightarrow \frac{\alpha + 1}{2} \left[F \left(r_0 \sqrt{\frac{2}{\alpha + 1}} \right) - 1 \right] \quad (93)$$

As $\sigma, \lambda \rightarrow -1$ along $(\sigma + 1) = \alpha(\lambda + 1)$;

$$G \rightarrow -\frac{\alpha + 1}{2} \left[F \left(r_0 \sqrt{\frac{2}{\alpha + 1}} \right) + 1 \right]$$

As for $G(T(\lambda), \sigma, r_0)$ we have, in addition to the asymptotic formulas in the table:

$$I_B' \text{ transforms into } = 1 + \frac{p}{\sigma + 1}.$$

The intersection of $(I_n)_T$ or $(I_n')_T$ with $\sigma = 0$ is given by

$$\lambda_{1,2} = p \pm \sqrt{1 - \frac{j_n^2}{r_0^2}}.$$

APPENDIX II. THE FIELD COMPONENTS

The field components are given here for the ferrite and for the plasma. They are normalized in such a way that E_z takes a simple form. It should be noted that the λ 's appearing in these equations are those defined in Secs. 4.11 and 4.2 for the ferrite and plasma respectively and have a different significance in the two cases.

We write

$$A_1(r) = \frac{J_n(\chi_1 r)}{J_n(\chi_1 r_0)} \quad A_2(r) = \frac{J_n(\chi_2 r)}{J_n(\chi_2 r_0)}$$

Then, for the ferrite,

$$E_z = [A_1(r) - A_2(r)]e^{jn\theta},$$

$$E_r = -j\beta \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \left\{ F_n(\chi_1 r) - \frac{n}{\lambda_1} \right\} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \left\{ F_n(\chi_2 r) - \frac{n}{\lambda_2} \right\} \right] e^{jn\theta},$$

$$E_\theta = -\beta \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \left\{ \frac{F_n(\chi_1 r)}{\lambda_1} - n \right\} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \left\{ \frac{F_n(\chi_2 r)}{\lambda_2} - n \right\} \right] e^{jn\theta},$$

$$H_z = j\beta \left[\frac{1}{\lambda_1} A_1(r) - \frac{1}{\lambda_2} A_2(r) \right] e^{jn\theta},$$

$$H_r = - \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \left\{ n - \frac{1 - \chi_1^2}{\lambda_1} F_n(\chi_1 r) \right\} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \left\{ n - \frac{1 - \chi_2^2}{\lambda_2} F_n(\chi_2 r) \right\} \right] e^{jn\theta}, \text{ and}$$

$$H_\theta = -j \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \left\{ F_n(\chi_1 r) - \frac{n}{\lambda_1} (1 - \chi_1^2) \right\} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \left\{ F_n(\chi_2 r) - \frac{n}{\lambda_2} (1 - \chi_2^2) \right\} \right] e^{jn\theta}$$

and, for the plasma,

$$E_z = [A_1(r) - A_2(r)]e^{jn\theta},$$

$$E_r = -\frac{j}{\beta} \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \{ (1 - \chi_1^2) F_n(\chi_1 r) + n\lambda_1 \} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \{ (1 - \chi_2^2) F_n(\chi_2 r) + n\lambda_2 \} \right] e^{jn\theta},$$

$$E_{\theta} = \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \{ \lambda_1 F_n(\chi_1 r) + n(1 - \chi_1^2) \} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \{ \lambda_2 F_n(\chi_2 r) + n(1 - \chi_2^2) \} \right] e^{jn\theta},$$

$$H_z = -\frac{j}{\beta} [\lambda_1 A_1(r) - \lambda_2 A_2(r)] e^{jn\theta},$$

$$H_r = -\left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \{ \lambda_1 F_n(\chi_1 r) + n \} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \{ \lambda_2 F_n(\chi_2 r) + n \} \right] e^{jn\theta},$$

and

$$H_{\theta} = -j \left[\frac{A_1(r)}{r} \frac{1}{\chi_1^2} \{ F_n(\chi_1 r) + n\lambda_1 \} - \frac{A_2(r)}{r} \frac{1}{\chi_2^2} \{ F_n(\chi_2 r) + n\lambda_2 \} \right] e^{jn\theta}$$

REFERENCES

1. Goldstein, L., Lampert M. A., and Heney, J. F., Phys. Rev., **82**, p. 956, 1951.
2. Polder, D., Phil. Mag., **40**, p. 99, 1949.
3. Hogan, C. L., B.S.T.J., **31**, p. 1, 1952.
4. Suhl H., and Walker, L. R., Phys. Rev., **86**, p. 122, 1952.
5. Kales, M. L., J. Appl. Phys., **24**, p. 604, 1953.
6. Gamo, H. J., J. Phys. Soc. Japan, **8**, pp. 176-182, 1953.
7. Cook, J. S., R. Compfner and H. Suhl, Letter to the Editor, I.R.E. Proc., to be published.

Coupled Wave Theory and Waveguide Applications

By S. E. MILLER

(Manuscript received February 2, 1954)

Some theory describing the behavior of two coupled waves is presented, and it is shown that this theory applies to coupled transmission lines. A loose-coupling theory, applicable when very little power is transferred between the coupled waves, shows how to taper the coupling distribution to minimize the length of the coupling region. A tight-coupling theory, applicable when the coupling is uniform along the direction of wave propagation, shows that a periodic exchange of energy between coupled waves takes place provided that the attenuation and phase constants (α and β respectively) are both equal, or provided that the phase constants are equal and the difference between the attenuation constants ($\alpha_1 - \alpha_2$) is small compared to the coefficient of coupling c . Either $(\alpha_1 - \alpha_2)/c$ or $(\beta_1 - \beta_2)/c$ being large compared to unity is sufficient to prevent appreciable energy exchange between the coupled waves. Experimental work has confirmed the theory. Applications include highly efficient pure-mode transducers in multi-mode systems, and frequency-selective filters.

INTRODUCTION

This paper describes some theoretical relations in coupled transmission lines, and the use of coupled lines as circuit elements. In order to illustrate the points of interest in the theoretical material, several applications will be stated first. Detailed discussion of experimental models will be given after the theoretical sections.

The theory of coupled transmission lines may be used to determine many properties of a multi-mode transmission system in which there is distributed coupling between modes. In round pipe, for example, the individual modes of propagation can be considered as separate transmission lines which in the perfect waveguide are completely independent. Geometric imperfections in the waveguide, if distributed over many wavelengths, cause a transfer of power between modes which in general

form is predicted by coupled transmission line theory. As a consequence, analysis of the mode-conversion effects associated with circular-electric-wave transmission in commercial round pipe has been aided materially by applying the coupled-transmission-line concept.¹ In another problem, the transmission of the circular-electric waves through bends,² the coupled-wave theory of subsequent sections has also provided valuable insight.

Coupled transmission lines can be employed as circuit elements to exchange power between one mode of a multi-mode line and a designated mode of another transmission line. Consider Fig. 1, which shows a rectangular waveguide having entries 1 and 2 coupled through a series of apertures to a parallel round waveguide having entries 3 and 4. The rectangular guide may be made single mode for convenience, and for the configuration shown may be made to couple to any TE mode of the round guide. Input power at entry 1 may be transferred in whole or in part to the selected mode at entry 4, the remaining portion of the power appearing at entry 2. Very little power in any mode will appear at entry 3 for excitation at 1, and very little power in undesired modes will appear at entry 4. Thus the structure has the hybrid property in addition to being mode selective. A matched impedance is presented at all entries to all modes over a very broad frequency band.

Recently, coupled transmission lines have found use as input and output circuits for travelling-wave tubes. In this instance a helical input (or output) line was electromagnetically coupled to the travelling-wave-tube helix, with conditions adjusted for complete energy transfer between the helices. The result is an input-output circuit requiring no metallic connection to the tube helix and requiring no connection through

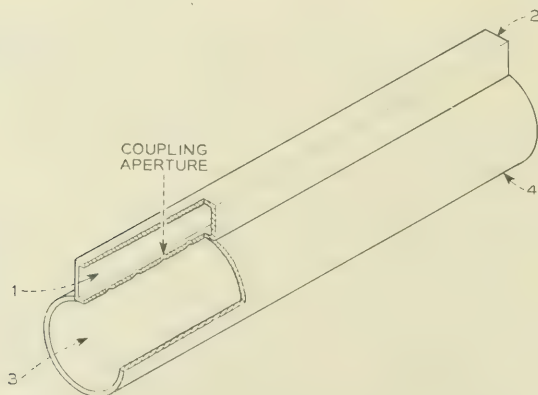


Fig. 1 — Coupled transmission line transducer.

the vacuum seal. R. Kompfner conceived this form of connection to travelling-wave tubes while working with the Admiralty in England, and demonstrated the usefulness of the idea here at the Laboratories. Similar work was done by the group at the Electronics Research Laboratory at Stanford University, and was described by S. T. Kaisel at the August, 1953, West Coast I.R.E. Convention. Both groups requested pre-publication copies of this paper for use in their research.

LOOSE COUPLING THEORY

On the assumption that negligible power is abstracted from the driven line of two coupled transmission lines, the magnitude and mode content of the forward and backward waves in the side line may be written. With reference to Fig. 2, there is assumed coupling between two uniform lines in the interval $-L/2$ to $+L/2$ along the axis of propagation, and no coupling elsewhere. On the basis of loose coupling a normalized voltage wave on line 2 may be written

$$E_2 = 1.0\epsilon^{-i(2\pi/\lambda_2)(x+L/2)}, \quad (1)$$

in which the phase reference is taken as $x = -L/2$. The forward current I_f in the side line at the point $x = L/2$ is

$$I_f = KFM \int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x) \epsilon^{i2\pi(1/\lambda_1 - 1/\lambda_2)x} dx, \quad (2)$$

where

$$F = \frac{\epsilon^{-i\pi L(1/\lambda_1 + 1/\lambda_2)}}{Z_{10}}.$$

$\phi(x)$ = a coupling function. More precisely, $1/\phi(x)$ is the ratio of the voltage on line 2, $E_2(x)$, to the equivalent voltage generator in series with line 1 at x .

K = fraction of the transferred current which travels in the forward direction.

M = the transfer constant for the various modes which can propagate, relative to the mode for which $\phi(x)$ is defined. The backward current I_b at the point $x = -L/2$ is

$$I_b = (1 - K)FM \int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x) \epsilon^{-i2\pi(1/\lambda_1 + 1/\lambda_2)x} dx. \quad (3)$$

If the coupling mechanism is non-directive (sending equal waves forward and backward) and has the same value for all modes, then $K = \frac{1}{2}$ and $M = 1.0$. For simplicity these values are assumed in writing the remainder of the expressions. However, the theory is applicable if the coupling mechanism is mode selective and/or directive provided that these properties do not change over the length of the coupling interval.

The mode discriminating property of the coupled lines is the ratio of the forward current for $\lambda_1 = \lambda_2$ to the forward current for $\lambda_1 \neq \lambda_2$. This ratio is

$$\text{Discrimination} = \left| \frac{I_f(\lambda_1 = \lambda_2)}{I_f(\lambda_1 \neq \lambda_2)} \right| = \frac{\int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x) dx}{\int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x) e^{i\frac{2\theta}{L}x} dx}, \quad (4)$$

where $\theta = \pi L(1/\lambda_1 - 1/\lambda_2) = L(\beta_1 - \beta_2)/2$ and the β 's are the phase constants of the two transmission lines.

The directivity of the coupling arrangement is defined as the ratio of the forward current for $\lambda_1 = \lambda_2$ to the backward current; this ratio is also given by equation (4) provided

$$\theta = -\pi L \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right) = -\frac{L}{2} (\beta_1 + \beta_2).$$

Thus, in the loose coupling case, the critical performance characteristics are given by the discrimination function, equation (4), for appropriate values of the parameter θ .

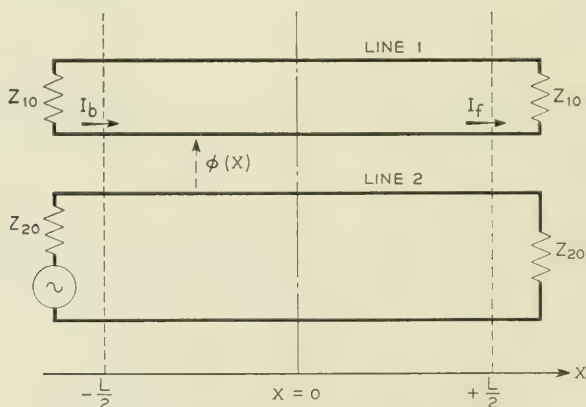


Fig. 2 — Schematic of coupled transmission lines.

A simplified example will illustrate the application of these relations. Suppose the coupling function $\phi(x)$ is constant in the interval $-L/2$ to $L/2$ and zero for other values of x . Then the discrimination function is, from (4)

$$\text{Uniform Coupling Discrimination} = \frac{\theta}{\sin \theta}. \quad (5)$$

Let us further assume, in the hypothetical example, that line 2 (Fig. 2) is a single-mode line having a guide wavelength λ_2 equal to $1.2\lambda_0$, and that line 1 is the three-mode line having guide wavelengths λ_1 , λ_2 , and λ_3 equal to $1.1\lambda_0$, $1.2\lambda_0$, and $1.3\lambda_0$ respectively. Assume the coupling length L equals $20\lambda_0$. For equal coupling to all modes in a differential unit of length, the relative current waves travelling in the forward direction in the three modes of line 1 are obtained from (4). For the ratio of the λ_2 forward current to the λ_1 forward current,

$$\theta = \pi 20\lambda_0 \left(\frac{1}{1.1\lambda_0} - \frac{1}{1.2\lambda_0} \right) = 1.52\pi$$

for which (5) gives a discrimination of about 13.5 db. For the ratio of the λ_2 forward current to the λ_3 forward current

$$\theta = \pi 20\lambda_0 \left(\frac{1}{1.2\lambda_0} - \frac{1}{1.3\lambda_0} \right) = 1.28\pi,$$

corresponding to a discrimination of about 14 db. For the ratio of the λ_2 forward current to the λ_2 backward current,

$$\theta = \pi 20\lambda_0 \left(\frac{1}{1.2\lambda_0} + \frac{1}{1.2\lambda_0} \right) = 33.3\pi,$$

corresponding to a discrimination of about 43 db. The backward currents in modes λ_1 and λ_3 can similarly be verified to be very small compared to the forward-travelling λ_2 current.

Thus, directivity and mode purity in a simplified case have been shown to be of the desired form.

It may be noted that the denominator of (4) is the Fourier transform of the coupling function $\phi(x)$. Since the numerator of (4) is independent of θ , the discrimination is maximized by minimizing the denominator. An analogous problem exists in the time versus frequency domain relations, and experience with the latter can be used to predict the discriminations to be expected using various coupling distributions.

In the simple example cited above, a length of coupling interval of $20\lambda_0$ yielded a discrimination between the desired versus undesired for-

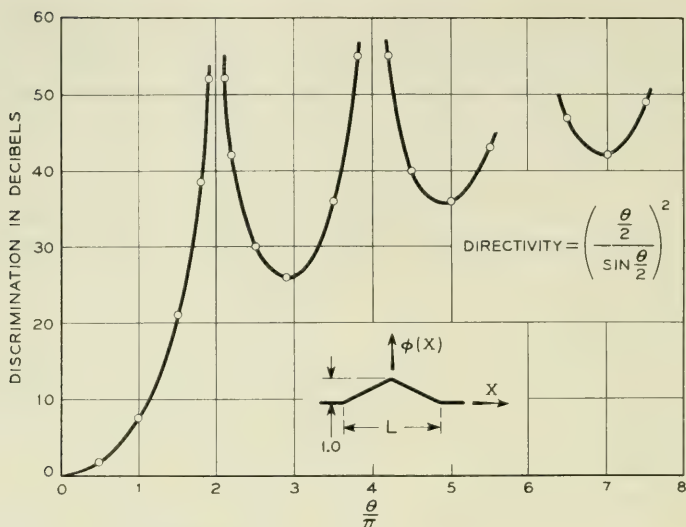


Fig. 3 — Discrimination versus θ/π for linear taper coupling.

ward wave components of about 13 db. How can this discrimination be improved? If the difference between the wavelengths of the desired and undesired wave types is increased, the value of θ is increased and greater discrimination results. In practical cases, however, there frequently is very little that can be done about the wavelength difference because it is inherent in a structure which is fixed by other considerations. By increasing the length of the coupling interval L the value of θ is also increased; in the case of uniform coupling, (5) shows that a value of θ/π equal to about 8 is required to get 30 db discrimination. In the above example this corresponds to L approximately equal to $125\lambda_0$. The latter coupling length is probably impractical, and is certainly inconvenient. The final alternative is to alter the distribution of coupling between the lines, and considerable can be done in this manner.

Suppose a linear taper of the strength of coupling is used, as sketched in Fig. 3. Then the discrimination becomes

$$\text{Linear Taper Discrimination} = \left(\frac{\theta/2}{\sin \theta/2} \right)^2. \quad (6)$$

which is plotted in Fig. 3. The first peak in discrimination occurs at θ/π equals two, compared to a value of θ/π equals one for the first peak using uniform coupling; however, for all values of θ/π greater than about 3, the linear taper provides superior discrimination. This illustrates a general trend; tapering the coupling distribution improves the discrimina-

tion for large θ/π values at the expense of an increased θ/π value for the first discrimination peak.

The first two lines of Fig. 4 give the discrimination functions for two forms of cosine taper; Fig. 5 shows a plot of the first function and Fig. 6 shows a plot of the second function for a particular case. These figures illustrate the importance of the slope at the ends of the coupling distribution. Comparing Fig. 5 with Fig. 3, Fig. 5 has a larger end-slope, shows a lower value of θ/π for the first peak in discrimination, but provides

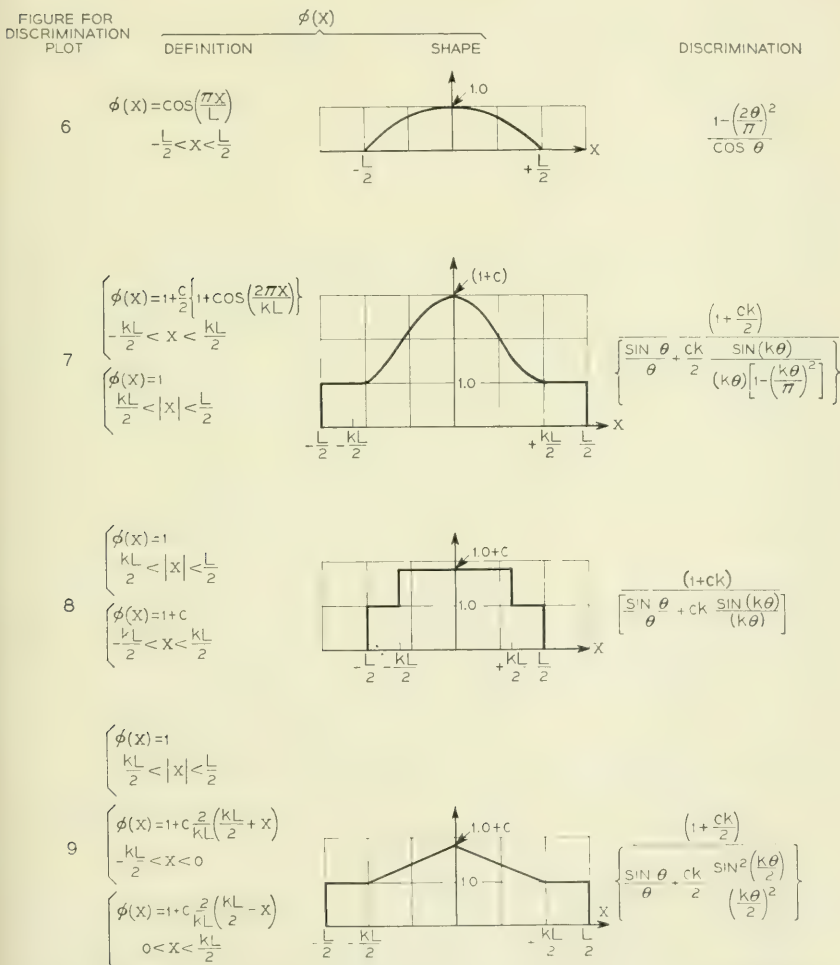


Fig. 4 — Discrimination functions corresponding to certain coupling distributions.

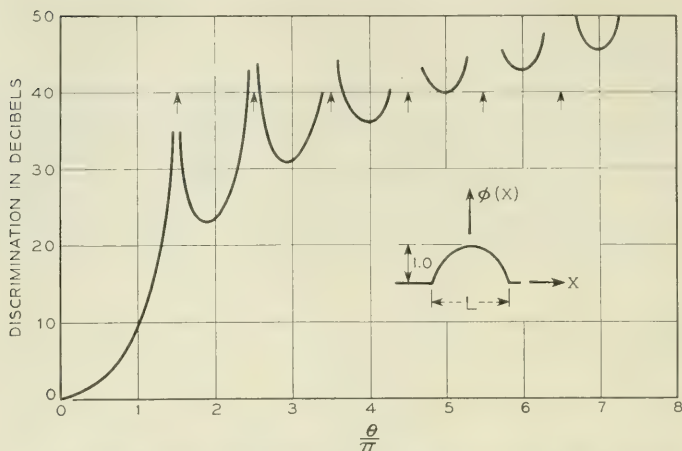


Fig. 5 — Discrimination versus θ/π for the cosine coupling distribution.

poorer discrimination at values of θ/π slightly above the first peak. In a similar way Fig. 6 shows better discrimination than Fig. 5.

Linear superposition of forward or backward currents may be employed to advantage when designing a coupling distribution. The second line of Fig. 4 gives the discrimination for a coupling function composed of a raised cosine plus uniform coupling. For a value of $c = 22.4$ and $k = 1$,

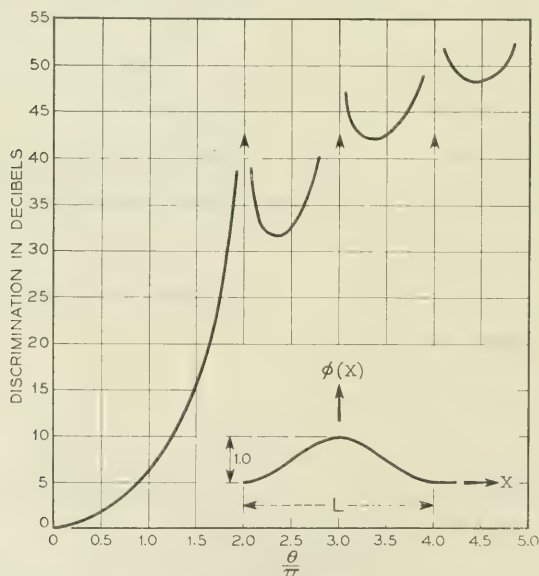


Fig. 6 — Distribution versus θ/π for the raised cosine coupling distribution.

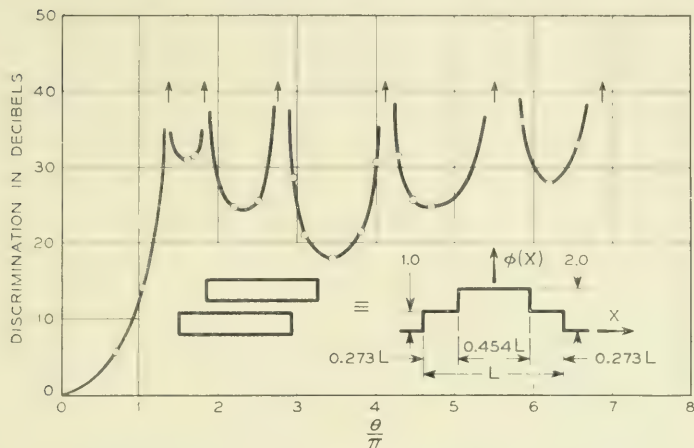


Fig. 7 — Discrimination versus θ/π for two uniform couplings superposed.

the discrimination is greater than 38 db for θ/π between 1.95 and 3.0, and is greater than 50 db for θ/π larger than 3. Below $\theta/\pi = 1.95$ the discrimination is similar to that shown in Fig. 6.

Linear superposition of two uniform coupling distributions yields a structure which is easy to fabricate and, in cases where the requirements are not too complex, may provide satisfactory discrimination. The third line of Fig. 4 gives the general relation, and Fig. 7 shows the discrimination plot for a case of interest. Discriminations on the order of 30 db are available in a broad region between θ/π equal to 1.3 to 2, an attractive abscissa value compared to the $\theta/\pi = 8$ required for simple uniform coupling.

Linear superposition of a linear taper and uniform coupling also yields a structure which is easy to fabricate, and the theoretical discrimination plot for an interesting set of conditions is shown in Fig. 8. High discriminations are provided over greater ranges of θ than for the case of two uniform coupling functions superposed.

The general relations involved in the superposition of coupling functions may be summarized as follows: Let $\phi_1(x)$, $\phi_2(x) \cdots \phi_n(x)$ be known coupling functions and let

$$\phi_T = \phi_1 + \phi_2 + \cdots \phi_n. \quad (7)$$

Let the maximum length of the coupling interval be L . Then, designating the transforms of ϕ_1 , $\phi_2 \cdots \phi_n$ by F_1 and $F_2 \cdots F_n$ respectively, where

$$F_n = \int_{-L/2}^{L/2} \phi_n(x) e^{i(2\theta/L)x} dx, \quad (8)$$

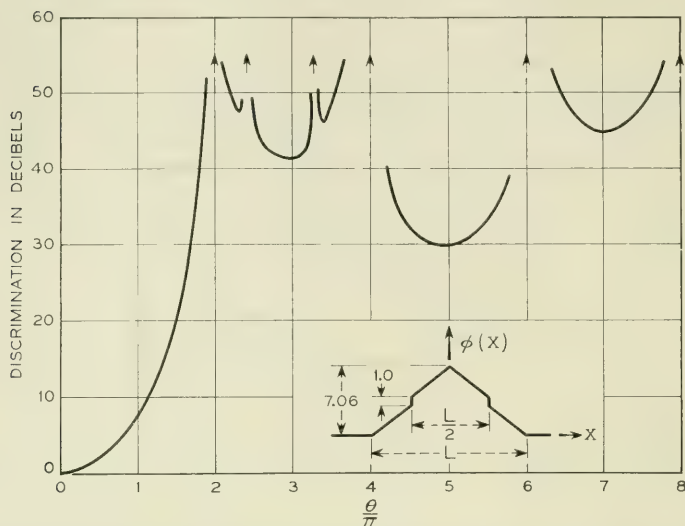


Fig. 8 — Discrimination versus θ/π for a linear taper and uniform coupling superposed.

and letting

$$F_T = F_1 + F_2 + \cdots F_n, \quad (9)$$

the discrimination function for the composite coupling distribution $\phi_T(x)$ is given by

$$\text{Discrimination} = \frac{F_T(\theta = 0)}{F_T}. \quad (10)$$

Another useful theoretical approach to the employment of multiple distributed coupling functions is illustrated in Fig. 9. The top sketch represents any coupling function $\phi_1(x)$. The lower sketch shows a new coupling function $\phi_2(x)$ formed by locating a $\phi_1(x)$ at $\pm d/2$ on the "x" axis. Using F_1 to denote the transform for $\phi_1(x)$, and F_2 to denote the transform corresponding to $\phi_2(x)$,

$$F_2 = 2F_1 \cos \theta', \quad (11)$$

wherein

$$\theta' = d \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right)$$

for the forward wave discrimination and

$$\theta' = d \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right)$$

for the directivity as defined earlier in connection with (4).

The discrimination function for the composite coupling function $\phi_2(x)$ is

$$\text{Discrimination} = \frac{F_2(\theta = 0, \theta' = 0)}{F_2} = \frac{F_1(\theta = 0)}{F_2} \frac{1}{\cos \theta'}. \quad (12)$$

The factor $1/\cos \theta'$ is the discrimination function associated with two point couplings, and the overall discrimination is the *product* of that discrimination and the discrimination associated with a single distributed coupling function $\phi_1(x)$. This line of thought may be extended to show that use of the same *distributed* coupling function in place of each point coupling in the multi-element distributions described in the following section results in multiplying the discrimination of the multi-element coupling function by the discrimination associated with the distributed coupling function.

In many cases of interest it is either inconvenient or impossible to use absolutely continuous coupling between transmission lines. In the waveguide case illustrated in Fig. 1, for example, a continuous slot cut in the common wall would not provide coupling of the distributed form due to a wave which would oscillate back and forth in the slot itself. We know, however, that the effects of the continuous coupling distribution can be

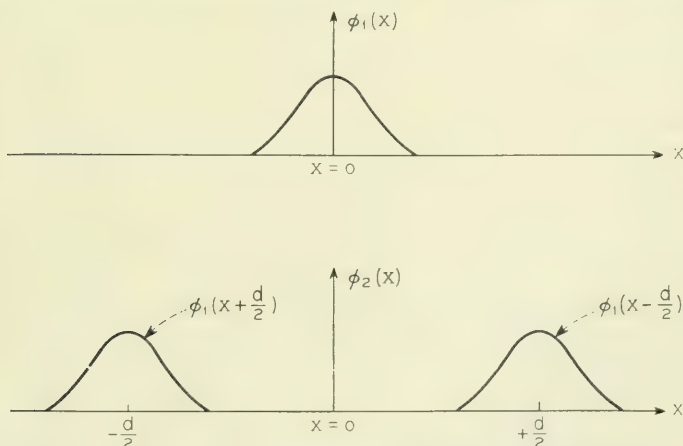


Fig. 9 — Schematic of multiple distributed coupling functions.

simulated closely by using closely spaced point couplings. In order to do this intelligently we need a theory for multi-element point couplings.

The most general symmetrical point coupling distribution for parallel coupled lines is illustrated in Fig. 10. The letters $a_0, a_1, a_2, \dots, a_n$ designate the strength of the couplings, and $d_1, d_2, \dots, L$ represents the spacings between them. The transform for the total coupling distribution is

$$F_T = \int_{-L/2}^{L/2} \phi_T \epsilon^{i(2\theta x/L)} dx. \quad (13)$$

$F_T = a_0 + 2a_1 \cos \gamma_1 + 2a_2 \cos \gamma_2 + 2a_3 \cos \gamma_3 + \dots + 2a_n \cos \theta$ in which

$$\gamma_k = \pi d_k \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right) \text{ or } \pi d_k \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right),$$

and

$$\theta = \pi L \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right) \text{ or } \pi L \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right),$$

depending on whether forward wave discrimination or directivity is required. The discrimination function is then

$$\text{Discrimination} = \frac{F_T(\gamma_k = 0, \theta = 0)}{F_T}. \quad (14)$$

Let us take as an example the familiar 1-3-3-1 binomial distribution of amplitudes for equally spaced couplings. In the terminology of equation (13), $a_0 = 0, a_1 = 3, a_2 = 1, a_k = 0$ for $k > 2, d_1 = L/3$, and $d_2 = L$. Then (14) yields

$$\text{Discrimination} = \frac{8}{6 \cos \theta/3 + 2 \cos \theta} = \frac{1}{\cos^3 \theta/3}. \quad (15)$$

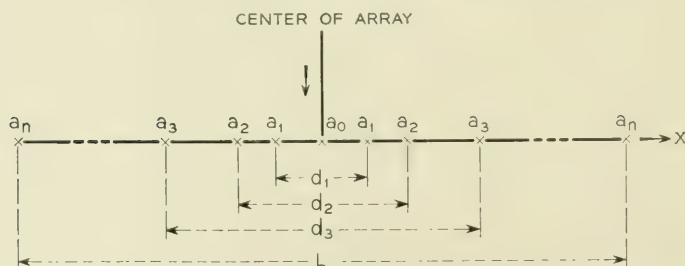


Fig. 10 — Schematic of point coupling distributions.

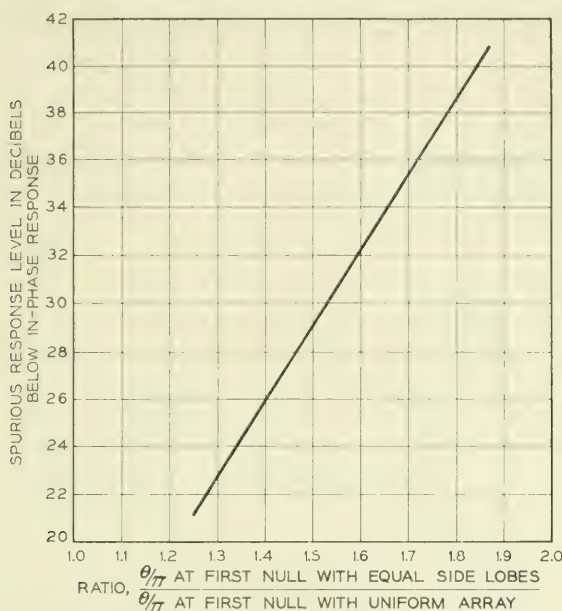


Fig. 11 — For all spurious mode responses down ordinate db, abscissa is the ratio of the required coupling length to the length required for constant amplitude coupling to produce a first null at the same value of $(1/\lambda_1 - 1/\lambda_2)$. (After C. L. Dolph, Reference 4).

which is the relation given by Mumford.³ The approach is perfectly general, and henceforth the coupling distribution only will be given with the understanding that the corresponding discrimination function can be obtained from (13) and (14).

For the case of tapered amplitudes and an even number of equally spaced couplings, (13) can be simplified to

$$F_T = 2a_1 \cos\left(\frac{\theta}{2n-1}\right) + 2a_2 \cos\left(\frac{3\theta}{2n-1}\right) + \cdots + 2a_n \cos \theta. \quad (16)$$

This case is of interest because a solution has been worked out for the analogous antenna problem to bring the spurious responses (the peaks of the side lobes in the antenna case, or the peaks of the undesired mode responses in the wave selector case) to the same level relative to the desired response. *This makes the total length of the coupling array a minimum for a given required degree of discrimination.* The solution⁴ includes specification of the Tchebysheff distribution of coupling strengths $a_1, a_2, a_3, \dots, a_n$ that are required to achieve various levels of spurious response, and the resulting increase in total array length required to place the first null

in undesired mode response at the same value of

$$\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \text{ or } \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)$$

as for uniform strength couplings equally spaced. Fig. 11 shows the latter relation, a very useful yardstick with which to evaluate the extra coupling length required by less ideal but more easily constructed coupling distributions.

An important practical question is "What is the smallest number of point couplings which will satisfy requirements in a given situation?", for it is time-consuming and expensive to fabricate the coupling holes or probes in some circumstances. The large range of possible mode conditions and discrimination requirements makes it difficult to give an answer in closed form, but the general restrictions involved may be stated. In the case of n equally spaced couplings (of any amplitude taper) the discrimination vanishes at $\theta/\pi = (n - 1)$. This is illustrated by the discrimination plot of Fig. 12.

Moreover, it is found that equally spaced couplings produce discriminations which are periodic in θ/π on the interval $(n - 1)$, and which are symmetrical about $\theta/\pi = (n - 1)/2$.

The implication of the discrimination zero at $\theta/\pi = (n - 1)$ is that a large number of point couplings are required to get good directivity and good forward wave discrimination. In the simple case cited above in which $L = 20\lambda_0$, the θ/π value for directivity was shown to be 33.3.

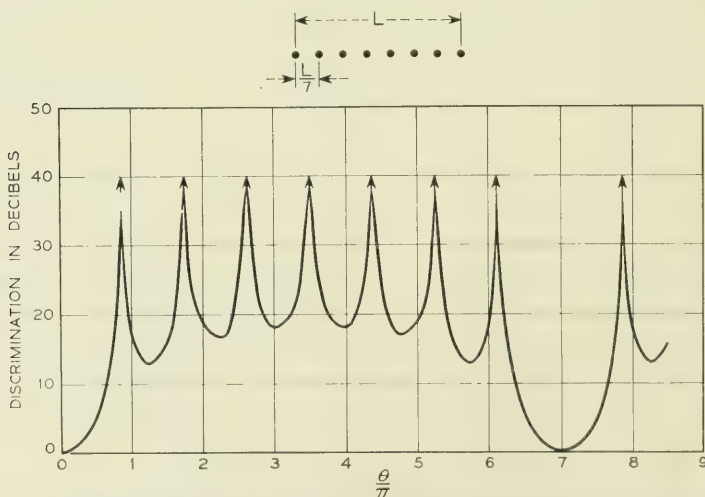


Fig. 12 — Discrimination for 8 equal-strength point couplings equally spaced.

Thus, something on the order of 50 or 60 equally spaced couplings might be needed.

Simulation of continuous coupling functions with equal strength couplings may be carried out as follows: the coupling amplitude versus distance plot may be divided along the distance axis into a number of intervals of equal area, and a point coupling placed at the center of each interval. The more efficient continuous coupling functions require more point couplings to get a good simulation in this manner. For example, the function of line 2, Fig. 4, with $c = 22.4$ and $k = 1$ has been simulated with 12 and 40 equal strength couplings (as described above) and the exact discrimination plotted using (13) and (14). The results are given in Figs. 13 and 14. The original continuous coupling function yields discriminations greater than 38 db for all values of θ/π greater than 2; the 40-point simulation approximates this well in the region of $\theta/\pi = 1.7$ to 4.5, but thereafter begins to fail. The 12-point simulation (Fig. 13) never matches the original but does best in the region of small θ/π .

It is more efficient to seek high discriminations by tapering the strength of equally spaced couplings than by tapering the spacing between equal strength couplings. However, when low discriminations are acceptable, the relative efficiency of tapering the spacing between con-

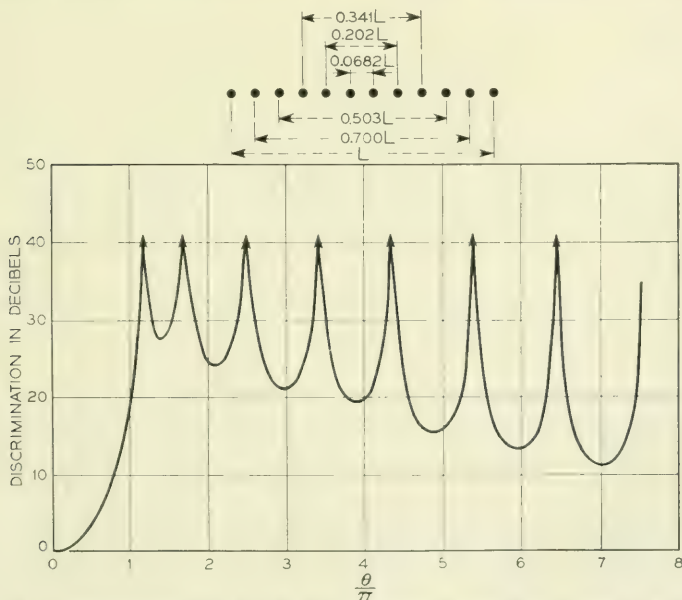


Fig. 13 — Discrimination for 12 equal-strength point couplings arranged to simulate the continuous distribution of Fig. 4, line 2.

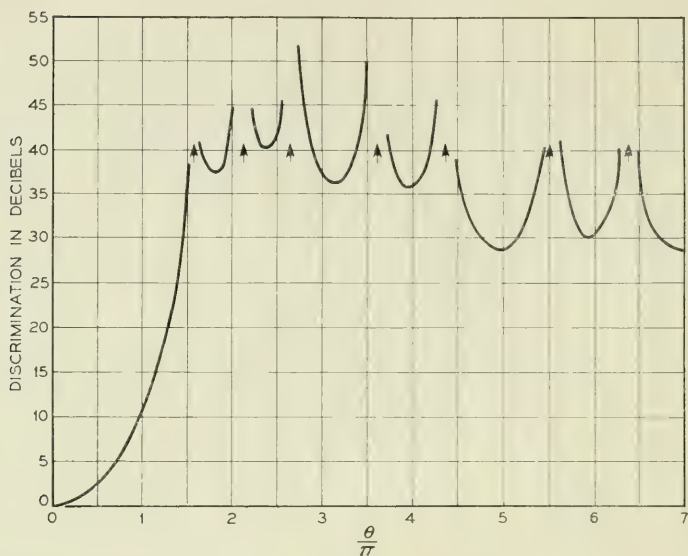


Fig. 14 — Discrimination for 40 equal-strength point couplings arranged to simulate the continuous distribution of Fig. 4, line 2.

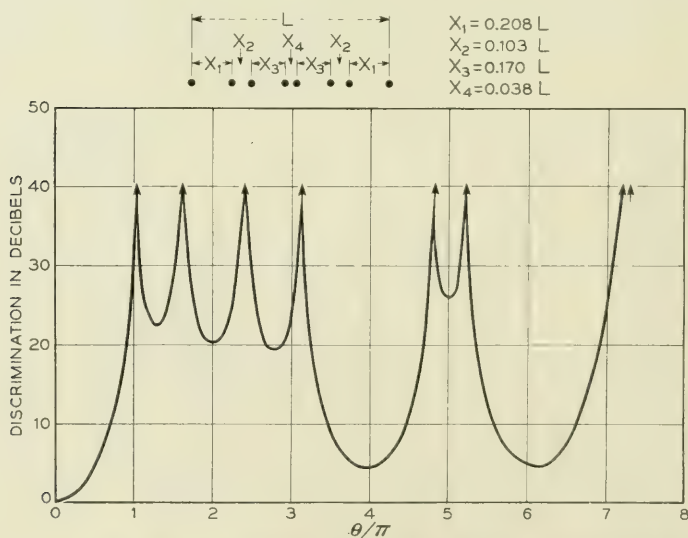


Fig. 15 — Discrimination function for 8 equal strength couplings arranged to maximize the bandwidth (θ/π range) for moderate discrimination.

stant strength couplings is much greater than when high discriminations are required. Fig. 15 shows a distribution which produces about 20 db discrimination from $\theta/\pi = 1$ to 3.25. Eight couplings arranged with the Tchebysheff amplitude taper for 20 db discrimination would produce that discrimination from $\theta/\pi = 1.05$ to 5.95.

It is possible to obtain directivity or mode discrimination at smaller θ/π values than made available with uniform coupling. This situation is analogous to the superdirectivity problem in antenna design, with similar results — the lobes of spurious response are increased. In particular, if the coupling near the ends of the third array of Fig. 4 is made larger than the coupling in the center region, making “ c ” a negative quantity, the first peak in discrimination occurs at θ/π less than one, and the first minimum in discrimination becomes less than 13 db.

By implication, emphasis has been placed on obtaining both mode discrimination and directivity simultaneously. However, by employing a relatively short coupling length it is apparent that the discrimination associated with

$$\theta = \pi L \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right)$$

may be kept small when the directivity associated with

$$\theta = \pi L \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right)$$

is in suitable range for good discrimination. Consequently, one can design a directional coupler with little mode discrimination. Conversely, when using a relatively small number of point couplings, the mode discrimination in the forward wave may be good when the directivity is poor.

TIGHT COUPLING THEORY*

We now consider the case in which a significant amount of power is taken from the driven transmission line by the line coupled to it. To simplify the problem the coupling is assumed uniform along the length

* An analysis of coupled transmission lines was given by W. J. Albersheim,¹¹ and the effects of coupling between waves on certain particular forms of transmission media were analyzed by Meyerhoff⁹ and Krasnushkin and Khokhlov.¹⁰ The treatment given here is intended to be more general and is believed to describe the effects of wave coupling under a greater variety of conditions.

axis. The space variation of the wave amplitude may be written

$$\frac{dE_1}{dx} = -(\Gamma_1 + k_{11})E_1 + k_{21}E_2, \quad (17)$$

and

$$\frac{dE_2}{dx} = k_{12}E_1 - (\Gamma_2 + k_{22})E_2, \quad (18)$$

in which k_{11} , k_{22} represent the reaction of the coupling mechanism on lines 1 and 2 respectively

k_{21} , k_{12} represent the transfer effects of the coupling mechanism

$\Gamma_{1,2}$ are the uncoupled propagation constants of line 1 and 2 respectively;

$E_{1,2}$ are the complex wave amplitudes on lines 1 and 2,

and are so chosen that $|E_1|^2$ and $|E_2|^2$ represent the power carried by lines 1 and 2 respectively at the input or output of the coupling region. The usual transmission-line equations are of this general form, except for second derivatives in place of the first. The first derivatives appear here because we deal only with the forward travelling waves, which the preceding section has shown are the only significant waves when small coupling per wave length is employed. Limiting our interest to the cases for which reciprocity holds and noting that there is always a transverse plane of symmetry midway between the ends of any pair of uniformly coupled lines, we may transform the wave amplitudes to make $k_{12} = k_{21} = k$. We may further simplify the equations without loss of essential generality by submerging the differences $(k_{11} - k)$ and $(k_{22} - k)$ into a modified propagation constant for lines 1 and 2 respectively, yielding

$$\frac{dE_1}{dx} = -(\gamma_1 + k)E_1 + kE_2, \quad (19)$$

and

$$\frac{dE_2}{dx} = kE_1 - (\gamma_2 + k)E_2, \quad (20)$$

in which

$$\gamma_1 = \Gamma_1 + k_{11} - k, \quad \text{and} \quad (20')$$

$$\gamma_2 = \Gamma_2 + k_{22} - k.$$

For some cases $k_{11} = k_{22} = k$ and for all cases of interest here γ_n differs very little from Γ_n since we are concerned only with loose coupling per wavelength.

The solution, for $E_1 = 1.0$ and $E_2 = 0$ at $x = 0$, is

$$E_1 = \left[\frac{1}{2} - \frac{(\gamma_1 - \gamma_2)}{2\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}} \right] \epsilon^{r_1 x} + \left[\frac{1}{2} + \frac{(\gamma_1 - \gamma_2)}{2\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}} \right] \epsilon^{r_2 x}, \quad (21)$$

and

$$E_2 = \frac{k}{\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}} \epsilon^{r_1 x} - \frac{k}{\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}} \epsilon^{r_2 x}, \quad (22)$$

where

$$r_1 = -\frac{1}{2}(2k + \gamma_1 + \gamma_2) + \frac{1}{2}\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}, \quad (23)$$

$$r_2 = -\frac{1}{2}(2k + \gamma_1 + \gamma_2) - \frac{1}{2}\sqrt{(\gamma_1 - \gamma_2)^2 + 4k^2}. \quad (24)$$

The nature of the coupling coefficient k is the first thing to investigate. Assume no dissipation in either the transmission line or in the coupling mechanism. Then it follows that for any value of x ,

$$|E_1|^2 + |E_2|^2 = \text{constant} \quad (25)$$

on the basis of energy conservation. It may be determined that (25) leads to the requirement that the coupling constant k be purely imaginary. This is a very important result. In all of the following discussion k is taken to be purely imaginary. Even where dissipation in the transmission lines themselves is important, it is still assumed that the coupling mechanism is non-dissipative.

The simplest case is $\gamma_1 = \gamma_2 = \gamma$, coupling between identical transmission lines. Then (21) and (22) reduce to

$$E_1 = \cos cx \epsilon^{-(ic+\gamma)x}, \quad (26)$$

and

$$E_2 = i \sin cx \epsilon^{-(ic+\gamma)x}, \quad (27)$$

where $k = ic$. The exponential of (26) and (27) shows that the coupling modifies the average phase constant, and that the attenuation in the driven line (E_1) is the same as in the uncoupled case for cx (coupling length times coupling strength) equal to $n\pi$ radians. The amplitude and phase variations due to the coupling are plotted in Fig. 16. Complete power transfer between lines takes place cyclically, with a period of $cx = \pi$, and with suitable choice of the product cx , an arbitrary division of power between the lines may be selected.

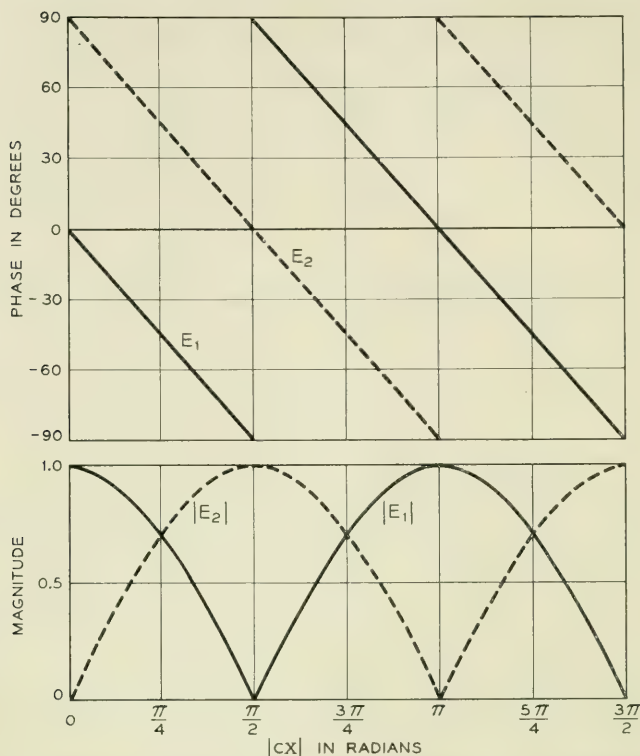


Fig. 16 — Wave amplitude and phase factors versus the integrated coupling strength $\underline{cx}$ for tightly coupled transmission lines having identical propagation constants.

Let us now assume that the phase constants of the two lines are unequal, but the attenuation constants are the same. Then

$$\begin{aligned}\alpha_1 &= \alpha_2 = \alpha, & \text{and} \\ (\gamma_1 - \gamma_2) &= i(\beta_1 - \beta_2),\end{aligned}\quad (28)$$

and equations (21) and (22) reduce to

$$E_1 = \epsilon^{-[\alpha + i(c + (\beta_1 + \beta_2)/2)]x} E_1^*, \quad (29)$$

where

$$\begin{aligned}E_1^* &= \cos \left[\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1} cx \right] \\ &- \frac{i(\beta_1 - \beta_2)}{2c} \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}} \sin \left[\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1} cx \right]\end{aligned}\quad (30)$$

$$E_2 = \epsilon^{-[\alpha + i(c + (\beta_1 + \beta_2)/2)]x} E_2^* \quad (31)$$

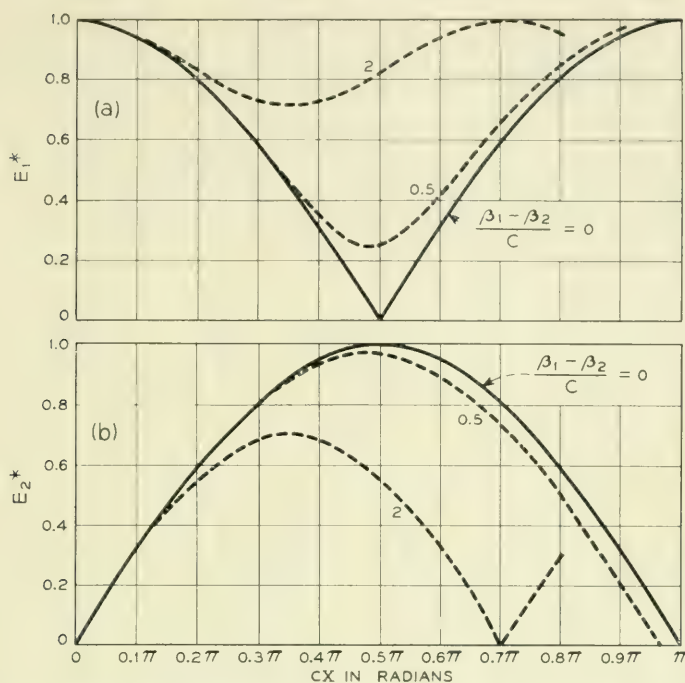


Fig. 17 — Wave amplitude and phase factors versus cx when the coupled lines have equal attenuation constants but unequal phase constants.

where

$$E_2^* = \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}} \frac{i}{\sin} \left[\sqrt{\frac{(\beta_1 + \beta_2)^2}{4c^2} + 1} cx \right] \quad (32)$$

The major effects of coupling in this case are represented by E_1^* and E_2^* , which are plotted in Fig. 17 for several values of $(\beta_1 - \beta_2)$. As $(\beta_1 - \beta_2)$ becomes different from zero, the maximum power transferred from the driven line to the undriven line decreases, and the period of the cyclical variation in amplitude is reduced. The latter period is the value of cx given by

$$\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1} cx = \pi \quad (33)$$

The driven and undriven-line wave amplitudes E_1^* and E_2^* at the maximum power transfer point, namely, at

$$\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1} cx = \frac{\pi}{2} \quad (34)$$

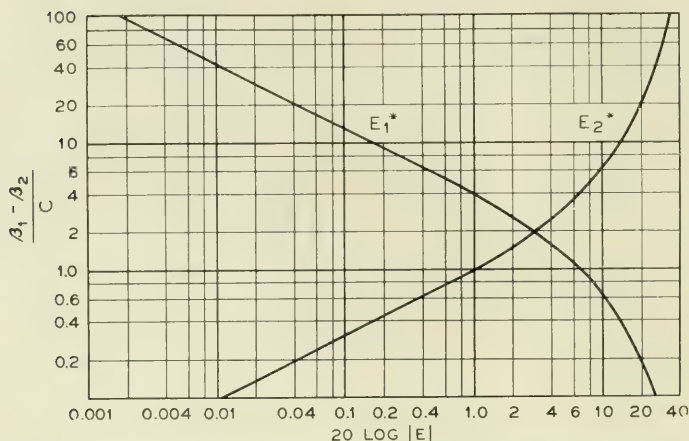


Fig. 18 — Wave amplitude factors at the maximum power transfer value of $\frac{cx}{\beta_1 - \beta_2}$ versus $(\beta_1 - \beta_2)/c$ when the coupled lines have equal attenuation constants.

are plotted in Fig. 18 as a function of the ratio $(\beta_1 - \beta_2)/c$. It is evident that this maximum energy transfer may be made very small for suitably large values of $(\beta_1 - \beta_2)/c$. The behavior of E_1^* and E_2^* as a function of coupling length x is shown with greater accuracy in the wide amplitude range of interest in Figures 19 and 20 respectively.

Consider now the case in which the coupled lines have identical phase constants, $\beta_1 = \beta_2 = \beta$, and unequal attenuation constants so that $(\gamma_1 - \gamma_2) = (\alpha_1 - \alpha_2)$. Then (21) and (22) reduce to

$$E_1 = e^{-[\alpha_1 + i(c+\beta)]x} \left\{ \left[\frac{1}{2} - \frac{(\alpha_1 - \alpha_2)}{2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}} \right] e^{[(\alpha_1 - \alpha_2)/2 + 1/2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}]x} \right. \quad (35)$$

$$\left. + \left[\frac{1}{2} + \frac{(\alpha_1 - \alpha_2)}{2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}} \right] e^{[(\alpha_1 - \alpha_2)/2 - 1/2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}]x} \right\},$$

$$E_1 = e^{-[\alpha_1 + i(c+\beta)]x} E_1^{**}, \quad (35')$$

and

$$E_2 = e^{-[\alpha_1 + i(c+\beta)]x} \frac{ic}{\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}} \left\{ e^{[(\alpha_1 - \alpha_2)/2 + 1/2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}]x} \right. \quad (36)$$

$$\left. - e^{[(\alpha_1 - \alpha_2)/2 - 1/2\sqrt{(\alpha_1 - \alpha_2)^2 - 4c^2}]x} \right\}, \quad \text{or}$$

$$E_2 = e^{-[\alpha_1 + i(c+\beta)]x} E_2^{**}. \quad (36')$$

The amplitude factors E_1^{**} and E_2^{**} have been defined in such a way as to reflect the principal effects of *attenuation difference* in the two lines; for the case in which the driven line attenuation constant α_1 is negligible, note that E_1^{**} and E_2^{**} contain all the amplitude variations of E_1 and E_2 respectively. In general, E_1^{**} and E_2^{**} are the ratios of the wave amplitudes actually present in lines 1 and 2 respectively to the wave amplitude which would exist in line 1 at the same value of x in the absence of coupling.

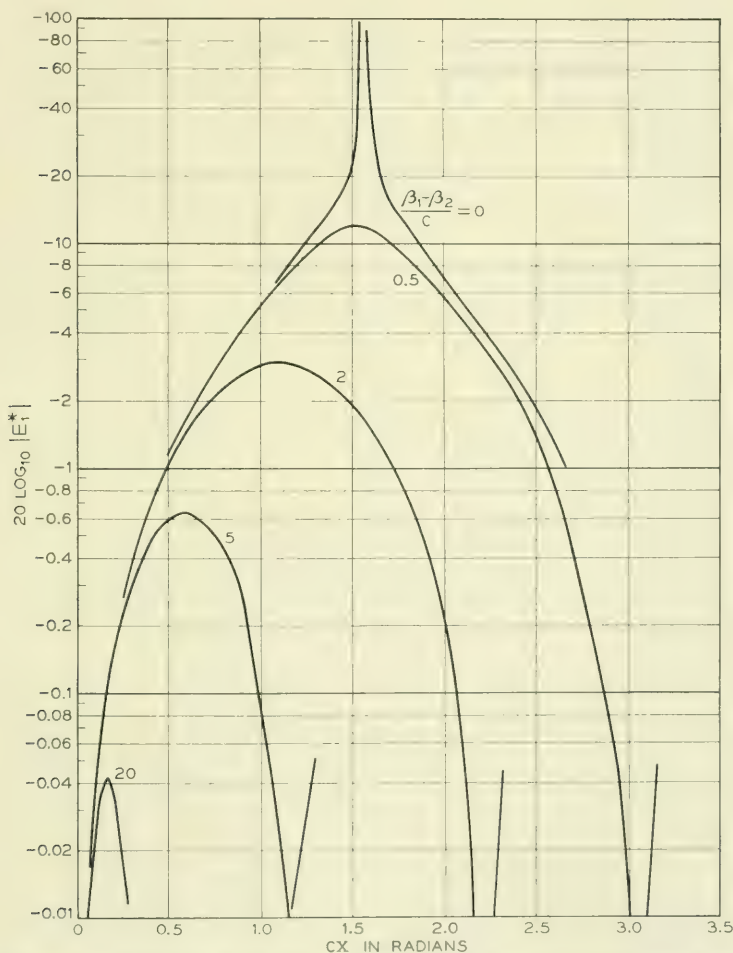


Fig. 19 — Driven line wave amplitude versus cx with unequal phase constants and equal attenuation constants. The curves are periodic for larger values of cx .

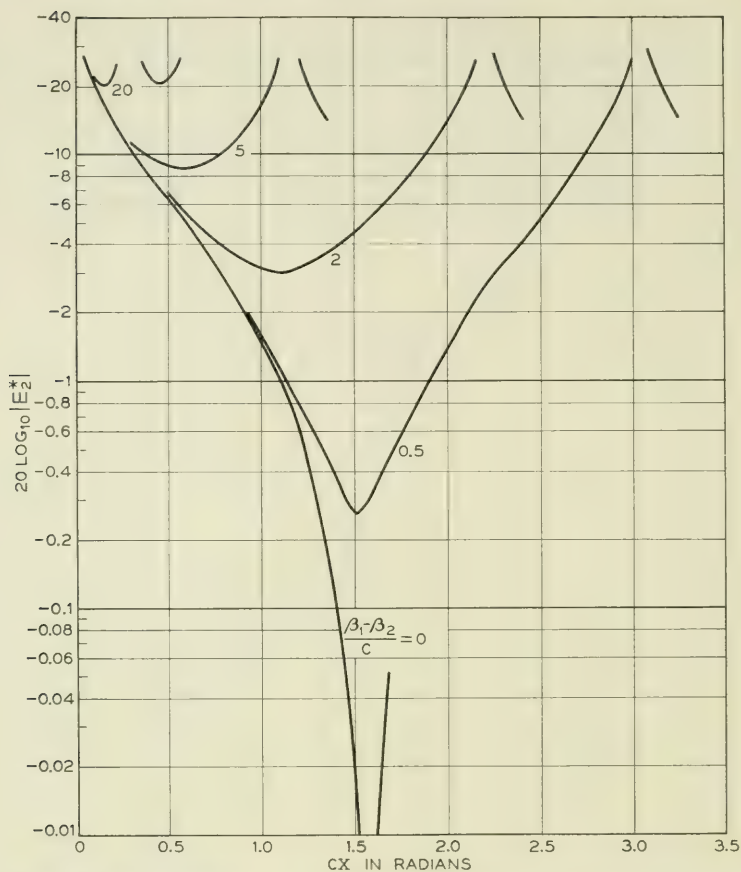


Fig. 20 — Undriven line wave amplitude versus cx with unequal phase constants and equal attenuation constants. The curves are periodic for larger values of cx .

We consider first the case of $(\alpha_1 - \alpha_2)$ negative, i.e., a lower attenuation constant in the driven line than in the undriven line. The effects of unequal attenuation constants may be illustrated at the integrated coupling strength $cx = \pi/2$ which, as Fig. 16 shows, results in complete transfer of power to the undriven line when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. Fig. 21 shows that the driven line wave amplitude E_1^{**} is very small when $(\alpha_1 - \alpha_2)/c$ is small, but is only $1\frac{1}{4}$ db below unity when $(\alpha_1 - \alpha_2)/c$ is about 55. Fig. 22 illustrates the way the undriven line wave amplitude E_2^{**} decreases as $(\alpha_1 - \alpha_2)/c$ increases.

For integrated coupling strengths less than $\pi/2$, the effects of unequal attenuation constants are not pronounced at small $(\alpha_1 - \alpha_2)/c$, but again for large $(\alpha_1 - \alpha_2)/c$, E_1^{**} approaches unity and E_2^{**} becomes small.

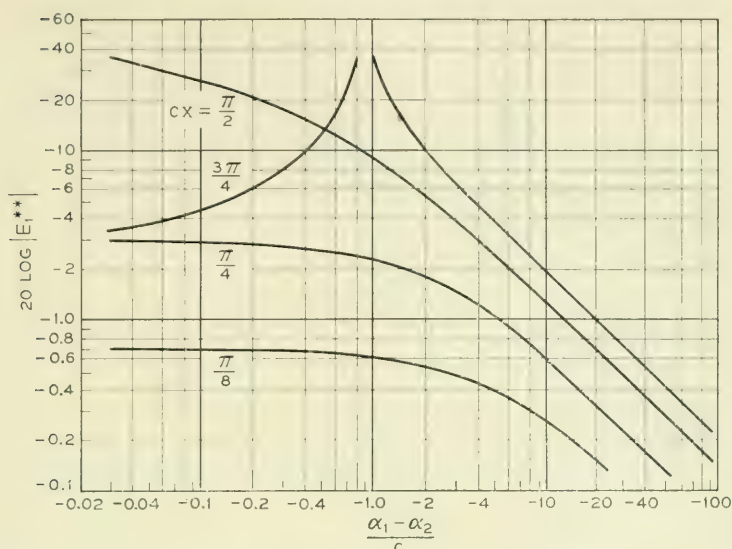


Fig. 21 — The effect of unequal attenuation constants on the driven line wave amplitude for equal phase constants, and $\underline{cx}$ constant. Negative $(\alpha_1 - \alpha_2)$ indicates that the undriven line has the larger attenuation constant.

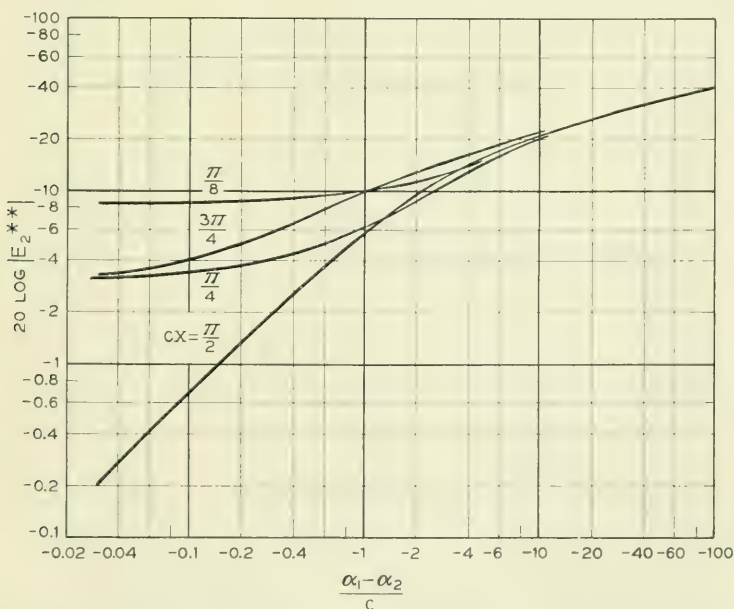


Fig. 22 — The effect of unequal attenuation constants on the undriven line wave amplitude for equal phase constants and $\underline{cx}$ constant.

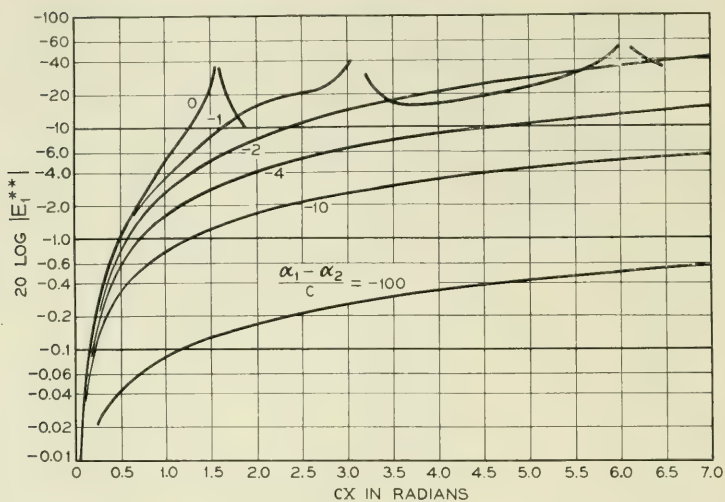


Fig. 23 — Driven line wave amplitude versus $\underline{cx}$ with equal phase constants and $(\alpha_1 - \alpha_2)/c$ as a parameter. The curve for $(\alpha_1 - \alpha_2)/c = 0$ is periodic.

For integrated coupling strengths greater than $\pi/2$ the effect of small values of $(\alpha_1 - \alpha_2)/c$ is to increase the loss to E_1^{**} , as shown by the curve for $\underline{cx} = 3\pi/4$ in Fig. 21. However, for sufficiently large values of $(\alpha_1 - \alpha_2)/c$ the loss to E_1^{**} is made small.

The variation in E_1^{**} and E_2^{**} as a function of coupling strength ($\underline{cx}$) is given in Figs. 23 and 24. The periodicity of E_1^{**} is removed for

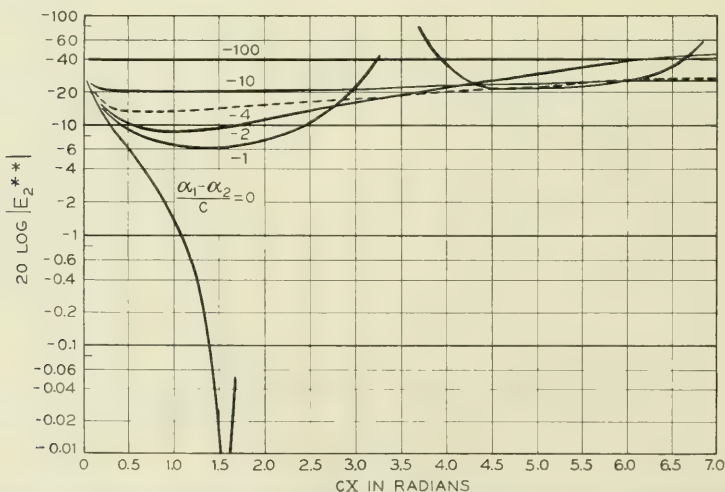


Fig. 24 — Undriven line wave amplitude versus $\underline{cx}$ with equal phase constants and $(\alpha_1 - \alpha_2)/c$ as a parameter. The curve for $(\alpha_1 - \alpha_2)/c = 0$ is periodic.

$(\alpha_1 - \alpha_2)/c$ as small as -1 , but a value as large as -10 or more is required in order to reduce the loss to E_1^{**} to a moderate value for large integrated coupling ($c\bar{x}$) values.

When $(\alpha_1 - \alpha_2)$ is positive, the attenuation constant for the undriven line is less than that for the driven line, and under these circumstances E_1^{**} can exceed unity. Physically this means that the power loss line is carrying the energy for a distance and returning it to the driven line at a more distant point. The curves of Fig. 25 and Fig. 26 show the variation of E_1^{**} and E_2^{**} versus positive $(\alpha_1 - \alpha_2)/c$ values, at fixed values of integrated coupling strength $c\bar{x}$. For $c\bar{x}$ equal to $\pi/4$, the driven line wave magnitude E_1^{**} decreases as the ratio $(\alpha_1 - \alpha_2)/c$ assumes small positive values and goes through a balanced type of null near $(\alpha_1 - \alpha_2)/c = 3.5$ (see Fig. 25). Again this is the resultant of the lower loss undriven wave carrying power for a distance and returning it to the driven wave in the proper phase to cause cancellation of the straight-through component of the driven wave. For $c\bar{x}$ between $\pi/4$ and $\pi/2$ the null would move from $(\alpha_1 - \alpha_2)/c$ near 3.5 toward $(\alpha_1 - \alpha_2)/c = 0$.

Figures 27 and 28 show the variation of E_1^{**} and E_2^{**} versus the integrated coupling strength $c\bar{x}$ at fixed values of $(\alpha_1 - \alpha_2)/c$. In these

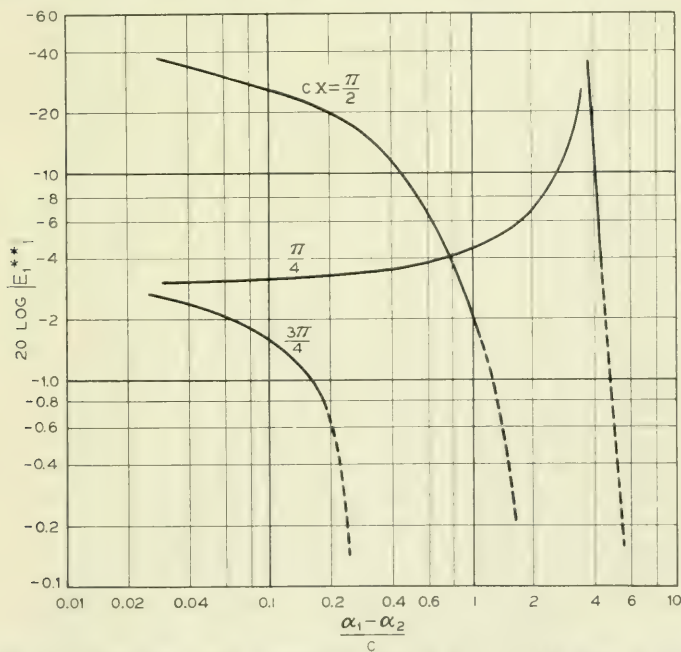


Fig. 25 — Driven line wave amplitude versus $(\alpha_1 - \alpha_2)/c$ with equal phase constants and $c\bar{x}$ constant. Positive $(\alpha_1 - \alpha_2)$ indicates the undriven line has the smaller attenuation constant.

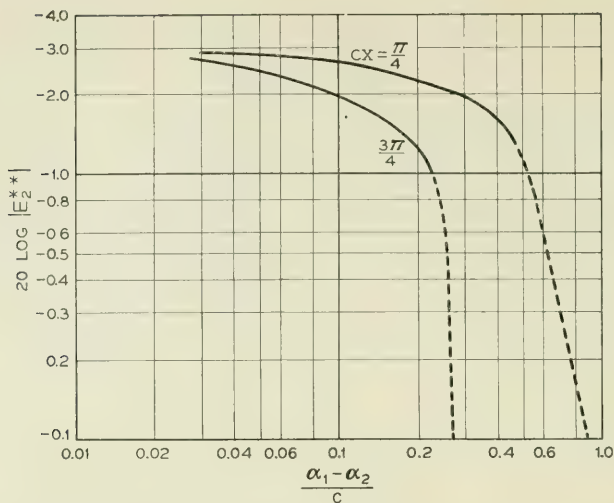


Fig. 26 — Undriven line wave amplitude versus $(\alpha_1 - \alpha_2)/c$ with equal phase constants and $\underline{cx}$ constant.

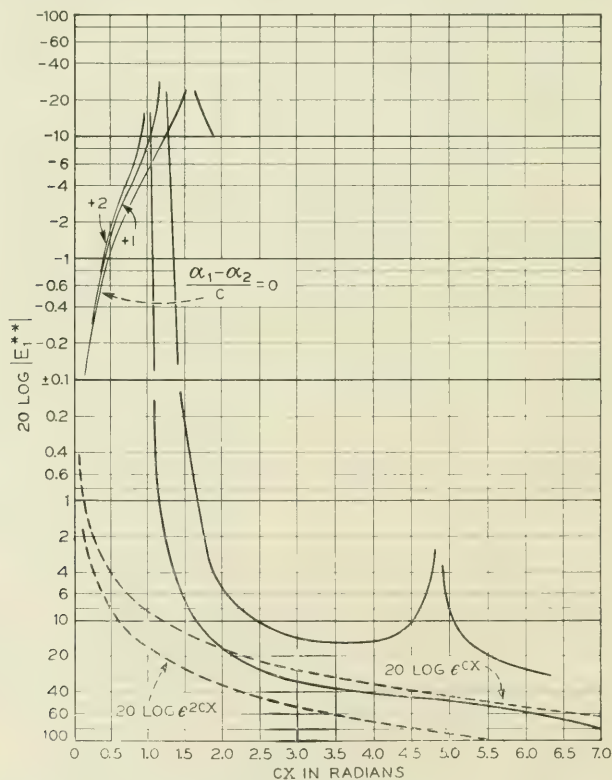


Fig. 27 — Driven line wave amplitude versus $\underline{cx}$, for equal phase constants; $(\alpha_1 - \alpha_2)/c$ as a parameter.

figures a double logarithmic scale is used on the ordinate to represent amplitude variations from 50 db below unity to amplitudes 50 db above unity. An arbitrary break in the scale has been made at ± 0.1 db which for practical purposes will be assumed to correspond to amplitudes of unity. With reference to Figure 27, small positive values of $(\alpha_1 - \alpha_2)/c$ move the first null in E_1^{**} from $c\bar{x} = \pi/2$ toward lower values of $c\bar{x}$. For abscissa values greater than $\pi/2$, E_1^{**} exceeds unity. For $(\alpha_1 - \alpha_2)/c = 1$, E_1^{**} again has a minimum in the vicinity of $c\bar{x} = 3\pi/2$ but this second null has disappeared for $(\alpha_1 - \alpha_2)/c = +2$ and presumably also for larger positive values. With reference to Figure 28, E_2^{**} grows at a more rapid rate as a function of $c\bar{x}$ when $(\alpha_1 - \alpha_2)/c$ takes on positive values. The null in the vicinity of $c\bar{x} = \pi$ is still present for $(\alpha_1 - \alpha_2)/c = 1$ but has disappeared at $(\alpha_1 - \alpha_2)/c = 2$. For $(\alpha_1 - \alpha_2)/c$ equal to $+2$ (and presumably for larger positive values) the undriven wave amplitude E_2^{**} is greater than E_1^{**} for $c\bar{x}$ larger than about 0.5.

The question comes to mind in connection with this case in which the

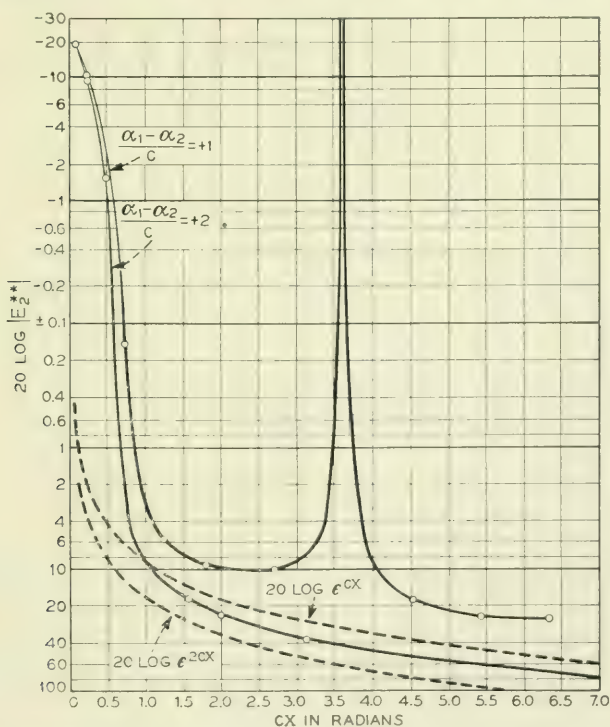


Fig. 28 - Undriven line wave amplitude versus $c\bar{x}$, for equal phase constants; $(\alpha_1 - \alpha_2)/c$ as a parameter.

undriven wave has a smaller attenuation coefficient than the driven wave, "How much less is the undriven line wave amplitude than would have existed at the same value of x if the same incident wave had been launched in the lower loss line and in the absence of coupling to the higher loss line?" This amplitude difference for the condition $(\alpha_1 - \alpha_2)/c = 1$ is represented in Fig. 28 by the *difference* between the curve for E_2^{**} and the curve labeled $20 \log e^{cx}$. Similarly, for the condition $(\alpha_1 - \alpha_2)/c = 2$, this amplitude difference is represented by the difference between the curve for E_2^{**} and the curve labeled $20 \log e^{2cx}$.

The general case of $\gamma_1 \neq \gamma_2$ is important both in interpreting undesired mode coupling effects in multi-mode systems as well as in evaluating

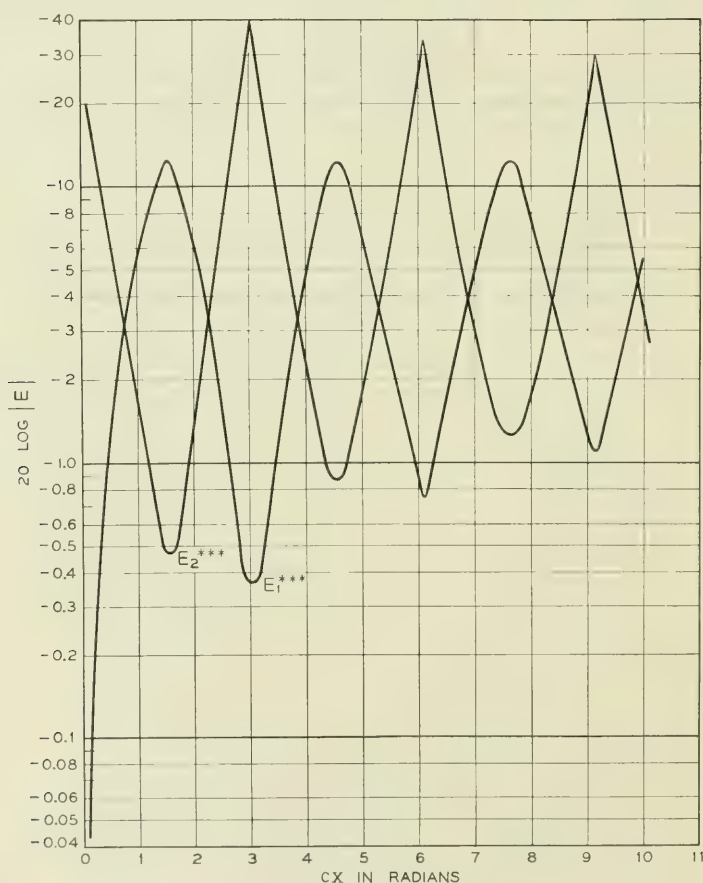


Fig. 29 — Driven and undriven line wave amplitudes versus $\underline{cx}$ with $(\alpha_1 - \alpha_2)/c = 0.03$ and $(\beta_1 - \beta_2)/c = 0.5$.

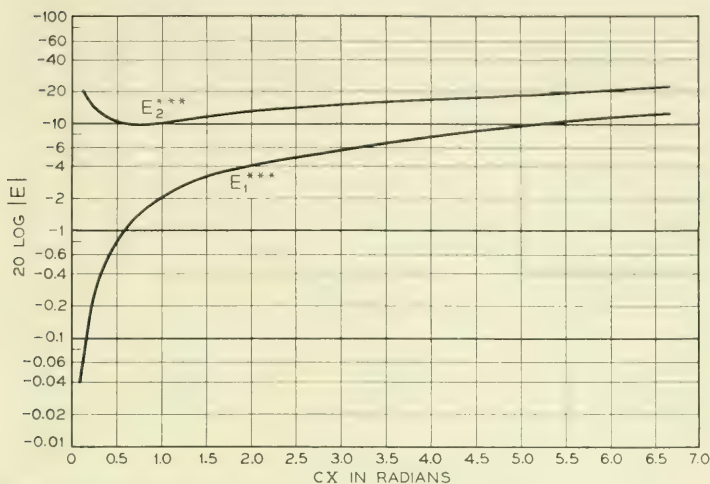


Fig. 30 — Driven and undriven line wave amplitudes versus $\underline{cx}$ with $(\alpha_1 - \alpha_2)/c = -2$ and $(\beta_1 - \beta_2)/c = 2$.

errors in construction of devices intended to produce $\gamma_1 = \gamma_2$. To facilitate discussion of this case we define

$$E_1 = E_1^{***} \epsilon^{-[\alpha_1 + i(c + (\beta_1 + \beta_2)/2)]x}, \quad (37)$$

and

$$E_2 = E_2^{***} \epsilon^{-[\alpha_1 + i(c + (\beta_1 + \beta_2)/2)]x}. \quad (38)$$

where E_1 and E_2 are defined by (21) through (24). The relation between E_1^{***} and E_1 (or E_2^{***} and E_2) is the same as described in connection with (35) and (36).

Small deviations from $\gamma_1 = \gamma_2$ are represented in Fig. 29, which shows E_1^{***} and E_2^{***} versus $\underline{cx}$ for $(\alpha_1 - \alpha_2)/c = -0.03$ and $(\beta_1 - \beta_2)/c = 0.5$. At $\underline{cx} = \pi/2$ radians, the first complete power transfer point in the $\gamma_1 = \gamma_2$ case, the above values correspond to a phase difference $(\beta_1 - \beta_2)x = \pi/4$ or 45° , and an attenuation difference $(\alpha_1 - \alpha_2)x = 0.03 \pi/2$ or 0.047 nepers (0.41 db) for the path length of the coupling distance. In the absence of the dissipation difference, but for the same difference in phase constants, Fig. 20 shows that E_2^* reaches a maximum at -0.26 db near $\underline{cx} = \pi/2$, whereas the value including the dissipation difference (Fig. 32) is -0.46 db. The latter two values differ by 0.2 db or one-half of $(\alpha_1 - \alpha_2)x$; when $(\alpha_1 - \alpha_2)/c$ is small compared to unity, this is a general result.

More sizeable deviations from $\gamma_1 = \gamma_2$ are represented in Fig. 30, which shows E_1^{***} and E_2^{***} versus $\underline{cx}$ for $(\alpha_1 - \alpha_2)/c = -2$ and $(\beta_1 - \beta_2)/c =$

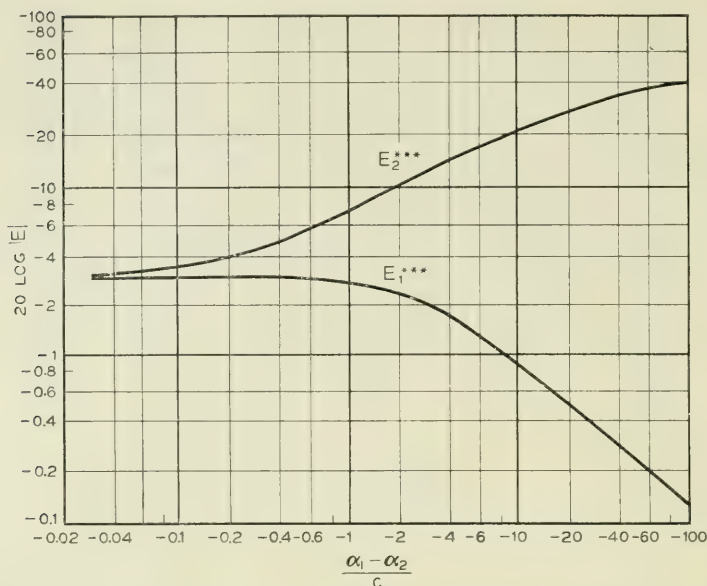


Fig. 31 — Driven and undriven line wave amplitudes versus $(\alpha_1 - \alpha_2)/c$ for $c x = \pi/2\sqrt{2}$ and $(\beta_1 - \beta_2)/c = 2$.

2. At $c x = \pi/2$, the phase difference is therefore π radians and the attenuation difference π nepers. The result is appreciable attenuation for E_1^{***} and only a moderate ratio of E_1^{***}/E_2^{***} .

Fig. 31 shows the way dissipation differences counteract the coupling forces when there is a phase constant difference $(\beta_1 - \beta_2)/c = 2$. This may be compared with Fig. 21 which represents the case of $(\beta_1 - \beta_2) = 0$. Very little change in E_1^{***} occurs until $(\alpha_1 - \alpha_2)/c$ exceeds $(\beta_1 - \beta_2)/c$; this is again a general result.

Finally, we may inquire as to how much power is dissipated in the system when attenuation constant differences are utilized to mitigate the effects of coupling. A measure of the power preserved is

$$|E_1^{***}|^2 + |E_2^{***}|^2$$

and this quantity is plotted in Fig. 32 for cases previously discussed in connection with Figs. 21 and 31. Either in the absence or presence of a phase constant difference, the attenuation constant difference shows a maximum effect in reducing the available power at $(\alpha_1 - \alpha_2)/c = 2$. This is probably a general result brought on by the factor

$$\sqrt{(\gamma_1 - \gamma_2)^2 - 4c^2}$$

found in the exponent of terms describing E_1 and E_2 .

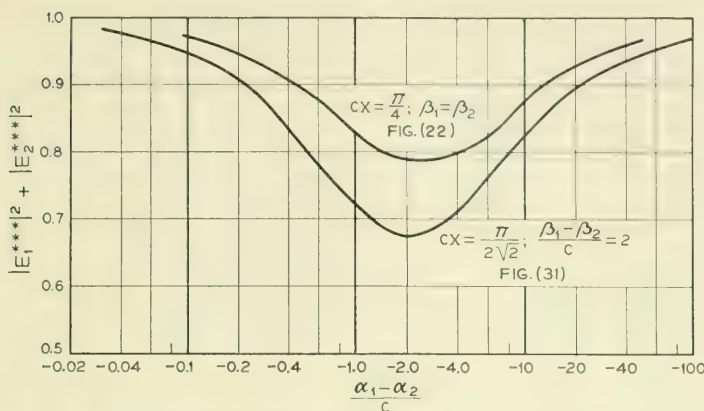


Fig. 32 — Available power versus $(\alpha_1 - \alpha_2) / c$ for several cases of interest.

TIGHT COUPLING EFFECTS OF MULTIPLE DISCRETE COUPLINGS

In practice it is convenient under some conditions to produce the desired coupling between transmission lines using multiple discrete couplings. It is then of interest to know the relation between the total power transferred and the number and strength of the individual couplings. It is the purpose of this section to state these relations.

We assume two transmission lines having identical propagation constants, with coupling units located at intervals along the lines as shown schematically in Fig. 33. A coupling unit may be a single point coupling, or an array of point couplings, but is always assumed to have the property of low reflection in the driven line and low back-wave transmission in the undriven line. If there are

n_1 couplings of magnitude α_1 ,

n_2 couplings of magnitude α_2 ,

and

n_k couplings of magnitude α_k

located along the lines in any order whatsoever, the wave amplitudes in

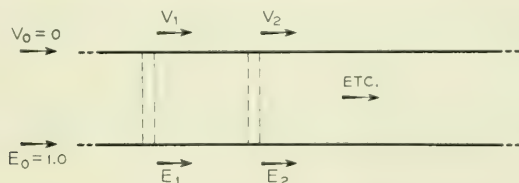


Fig. 33 — Schematic of transmission lines with multiple point couplings.

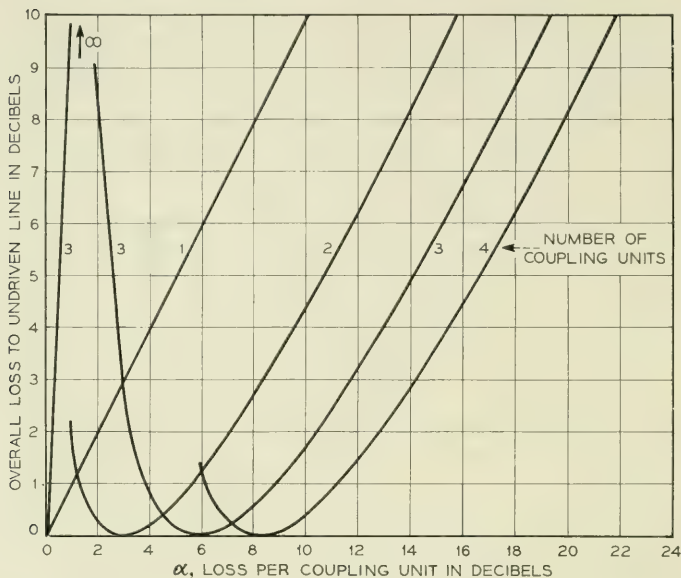


Fig. 34 — Overall loss to the undriven line versus loss per coupling unit, with the number of coupling units as a parameter.

the driven and undriven lines respectively are

$$E = \cos [n_1 \sin^{-1} a_1 + n_2 \sin^{-1} \alpha_2 + \cdots n_k \sin^{-1} \alpha_k], \quad (39)$$

and

$$V = \sin [n_1 \sin^{-1} \alpha_1 + n_2 \sin^{-1} \alpha_2 + \cdots n_k \sin^{-1} \alpha_k]. \quad (40)$$

These are amplitude factors due to coupling, and the normal attenuation effects in the uncoupled lines must be added separately. For complete power transfer we set the bracketed quantity of (39) and (40) equal to $\pi/2$, which gives the desired information about number and strength of point couplings. Other transfer losses may similarly be prescribed or determined.

For multiple coupling units of the same coupling strength, Fig. 34 shows the overall transfer loss to the undriven line versus loss per coupling units as a parameter. The shape of these curves from the complete transfer point toward higher losses is very nearly the same. Fig. 35 shows the loss per coupling unit versus number of coupling units, with overall transfer loss to the undriven line as a parameter.

SOME RESULTS OF EXPERIMENTS IN DOMINANT-MODE WAVEGUIDE

In a previous paper on dominant-mode waveguide directional couplers,⁵ complete power transfer between dominant-mode rectangular

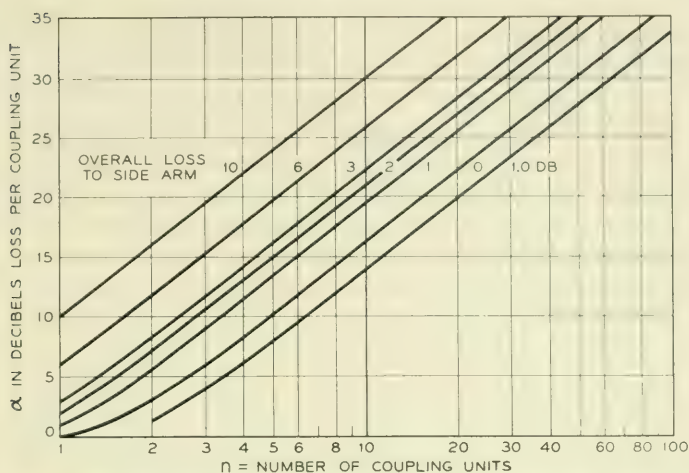


Fig. 35 — Loss per coupling unit versus number of coupling units, with the desired transfer loss as a parameter.

waveguides was shown to be possible in a coupling interval two wavelengths long, and very broad band directivity characteristics of a shape prescribed to meet given requirements were shown to be achievable.

The following paragraphs report on experiments which have been carried out with the objective of developing other useful devices and with the ancillary aim of verifying other predictions of the theory.

Experimental work was done to verify the cyclical nature of energy transfer between coupled lines, to determine the magnitude of losses which accompany such transfer in the waveguide case, and to determine desirable coupling distribution shapes in the tight coupling case. These experiments were carried out by R. W. Dawson in the 3.1 to 3.5 cm band using the 0.4" x 0.9" I.D. jig shown in Fig. 36, consisting of two wave-

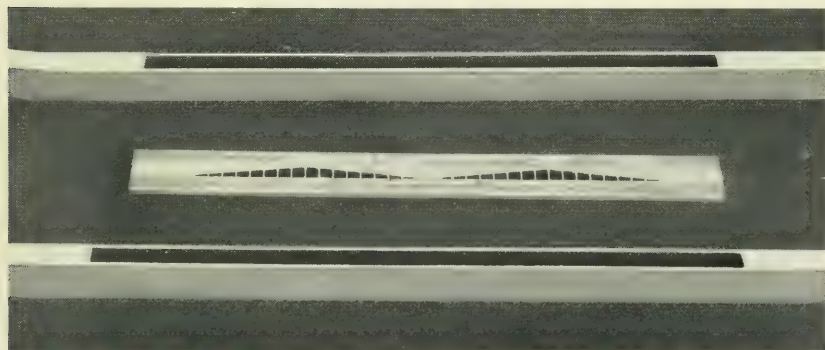


Fig. 36 — A 0.4" x 0.9" I.D. waveguide jig used for 3 cm coupled line experiments. The long waveguides on one side of the coupling insert were required to accommodate low-reflection terminations for directivity measurements.

guides having one wall cut away to accept a coupling insert. In one set of observations, the insertion loss in the driven-line and the transfer loss to the undriven line were recorded for a variable number of No. 22 copper wires dividing a coupling aperture $11\frac{1}{4}$ " long and linearly tapered from 0.030" height at the ends to 0.33" height at the center. The results are recorded in Fig. 37. At 102 holes, negligible power was abstracted from the driven line, and the transfer loss to the undriven line forward wave $|E_2|$ was about 18 db. Note that more coupling was observed at

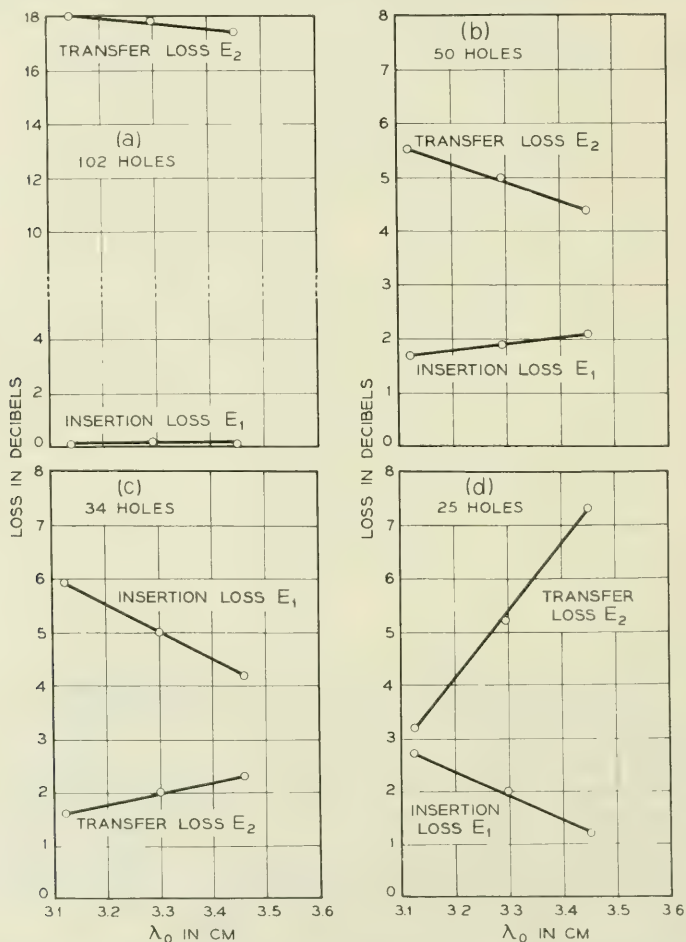


Fig. 37 — The transfer loss and the insertion loss versus frequency for the coupled waveguides of Fig. 36, showing the cyclical exchange of power as the coupling was increased by reducing the number of dividing wires in the fixed coupling aperture.

increasing wavelength values, a general result for small holes in the side wall. As the number of wires in the given aperture was reduced, markedly increased coupling resulted. This was due to the fact that the coupling loss per hole varied approximately as the fourth power of the hole dimension perpendicular to the electric vector, whereas the overall power loss varied only as the square of the number of holes in the loose coupling region. (Equations (39) and (40) describe the effects of number of coupling points more precisely.) At 50 holes, the transfer loss was about 5 db and the wave in the driven line was reduced by about 2 db; the slope of the $|E_2|$ versus λ_0 plot was the same as for 102 holes. At 34 holes, the transfer loss was about 2 db and the wave in the driven line was reduced by about 5 db; in this case, however, the undriven line wave loss increased with increasing λ_0 . Since coupling increases with increasing λ_0 we deduced that the total coupling was greater than required for complete power transfer and the bracketed expression of (39) and (40) was greater than $\pi/2$. On the diagram of Fig. 16, the presumed operating point was near $c\bar{x} = 2.2$ radians. At 25 holes, Fig. 37(d), the transfer loss was about 5 db and the wave in the driven line was reduced by about 2 db; as in the 34 hole case, the undriven line wave amplitude decreased with increasing λ_0 and hence with increasing coupling. Again the integrated coupling appeared to be in the region between $\pi/2$ and π . The driven line wave loss was headed for a low value at the long-wave end of Fig. 37(d), and it seems clear that periodic energy exchange is realized in practice.

The losses associated with this energy exchange may be inferred by comparing the total power output of the undriven and driven lines to the input power. Assuming that the forward waves in the driven and undriven lines contain all the output power, (i.e. neglecting reflection, back wave in the undriven line and waveguide losses) the following table gives the losses observed in the above described experiments:

Number of Holes	Coupling Mechanism Loss
	db
50	0.16
34	0.23
25	0.33

These losses may be due to circulating currents in the wires, in which case the loss would be expected to increase with increasing coupling.

Good agreement between the observed and theoretical directivities has been found in the loose coupling case,⁵ but when appreciable power is

abstracted from the driven line it is clear that the theory given above does not apply. Dawson has obtained experimental data of interest in this connection. For a $6\lambda_g$ long linear-taper aperture of the form given above (0.33" height at the center and 0.030" height at the ends), the loose coupling theory predicts directivities in excess of 45 db for the wavelength band 3.1 to 3.5 cm. When using sufficient number of wires to obtain 18 db transfer loss, directivities in the range 36 to 48 db were observed. The reason for the 36 db observation being lower than the 45 db theoretical value may be inaccuracy of fabrication (jig per Fig. 36) or inapplicability of the loose coupling theory at 18 db transfer loss. At 3 db transfer loss, the observed directivity of a similarly shaped but $5.5\lambda_g$ long coupling array is shown at the top of Fig. 38; again loose coupling theory predicts more than 45 db directivity. The reason the observed values are in the 24–33 db range rather than above 45 db is presumed to

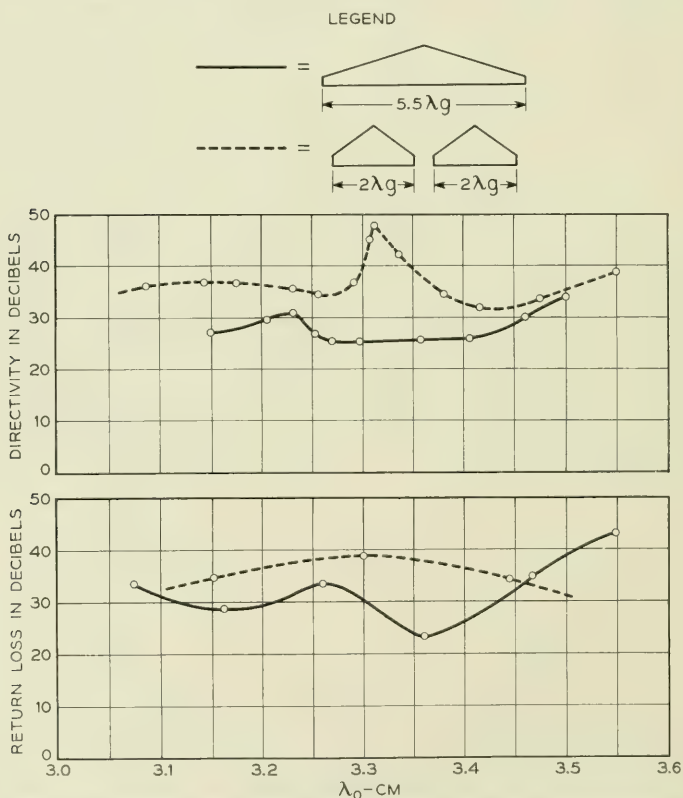


Fig. 38 — Directivity and return loss for two coupling distributions, each of which produced 3 db transfer loss.

TABLE I

Array	Transfer Loss		
	$\lambda_0 = 3.15$ cm	$\lambda_0 = 3.30$ cm	$\lambda_0 = 3.45$ cm
	db	db	db
Single $2\lambda_g$ array	3.2	3.0	2.8
Two cascaded arrays	4.3	4.2	4.0
Single $5.5\lambda_g$ array	3.5	2.9	2.4
Array	Straight Through Loss		
	$\lambda_0 = 3.15$ cm	$\lambda_0 = 3.30$ cm	$\lambda_0 = 3.45$ cm
	db	db	db
Single $2\lambda_g$ array	3.0	3.2	3.5
Two cascaded arrays	2.6	2.8	3.0
Single $5.5\lambda_g$ array	2.9	3.6	4.3

be inapplicability of the theory. Loose coupling theory predicts better than 35 db directivity over a broad frequency band at a coupling length of about $2\lambda_g$; therefore one might expect to obtain better overall results by using two cascaded arrays each about $2\lambda_g$ long, and each having a transfer loss of 8.4 db to get the 3 db net transfer loss. Observed directivities for such a coupling array are also given in the top of Fig. 38; in this case values in the 32–37 db region were obtained. The destructive interference associated with addition of backward wave components is more nearly of the form computed by loose coupling theory because the exciting wave is more nearly constant over the length of one of the arrays. The observed return loss at any one of the four waveguide entries, when the others are terminated, is given for the $5.5\lambda_g$ and cascaded $2\lambda_g$ coupling arrays at the bottom of Fig. 38. The cascaded $2\lambda_g$ combination is again superior to the single long taper. The characteristic of being inherently matched at all terminals makes the coupled-line type of 3 db hybrid attractive at the very high frequencies where lumped element matching becomes difficult if not impracticable.

Where space is at a premium, or where more constant transfer loss values are to be desired a shorter array composed of larger holes is attractive. A single linear taper of the shape outlined above and $2\lambda_g$ long was observed to have better than 22 db directivity and better than 25 db return loss over the 3.1 to 3.5 cm band. The observed loss values of the three coupling arrays discussed above are given in Table I. The coupling arrays composed of larger holes have less slope in the loss versus frequency characteristic for side-wall coupling.

SOME LOOSELY COUPLED TRANSDUCERS IN MULTIMODE WAVEGUIDE

In connection with research on low-loss circular-electric-wave transmission,¹ there developed a need for means with which to measure the power present in any one of the modes of a multi-mode round waveguide. In particular, it was known that the circular electric wave in round waveguide converts readily to the TM_{11} wave due to curvature of the line,² and a direct measurement of the effect was needed. The TM_{11} wave will not exist in the round waveguide without the presence of at least four other modes, and in the waveguide size used for the experiments five other modes could propagate. In designing a transducer for this application, therefore, it was necessary to evaluate the discrimination function, equation (4), with regard to mode discrimination between five different pairs of modes as well as to insure directivity. Moreover, the TM_{11} wave is degenerate with the circular electric wave TE_{01} , i.e., they have the same phase constant. Therefore, mode discrimination against TE_{01} could not be obtained through the phase difference effects described by (4). This discrimination was obtained using geometric balance in the individual coupling orifices, which were narrow slits on the center line of the wide side of the rectangular guide, as shown in Fig. 39. The shape of the coupling distribution employed was that described in connection with Fig. 14 except that 80 point couplings were used to simulate the raised-cosine coupling distribution (instead of 40 as in Fig. 14) in order to assure good directivity for the very long coupling length that was required. The round guide diameter was two inches, the rectangular guide width 0.820 inches, calculated to produce the same cut-off frequency in the rectangular guide as exists for the TM_{11} wave in the round waveguide. The coupling length was about 17 inches.

One simple method for evaluating the mode content of such a transducer is to measure the azimuthal distribution of electric field at the round guide wall using the radial probe technique described by M. Aronoff.⁷ If the power in a single mode is a great deal larger than the

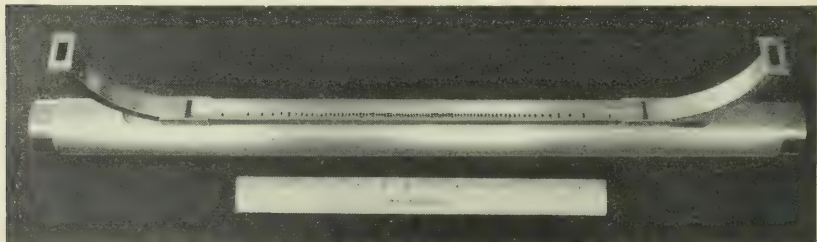


Fig. 39 — A TE_{10} to TM_{11} coupled wave transducer.

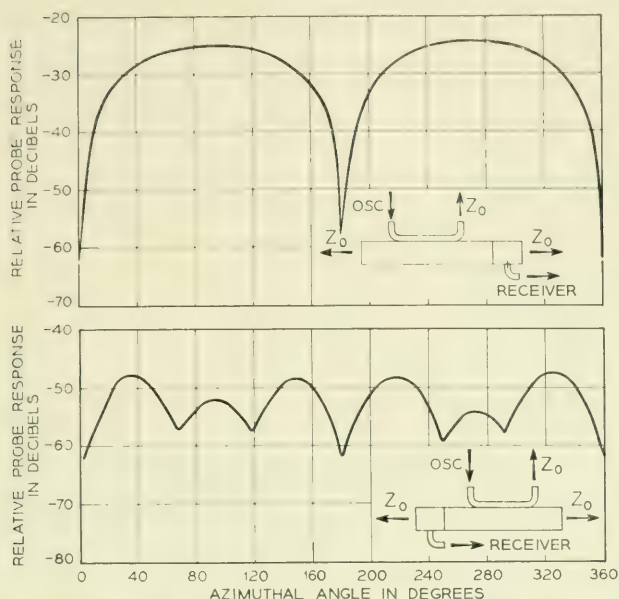


Fig. 40 — Distribution of radial electric field at the guide wall for the forward and backward waves of the transducer of Fig. 39.

power in any other mode of the multi-mode guide, the radial probe technique narrows down the possible mode types to a very few. Measurements of this type, recorded in Fig. 40, indicate that the forward wave has the radial electric field distribution to be expected for the TM_{11} wave. However, the forward wave might have the same radial field distribution at the wall and actually be the TE_{11} wave instead of TM_{11} . The TE_{11} wave is very simply generated from a dominant mode rectangular guide, by means of a long taper transition along the axis of propagation from the rectangular cross section to the circular cross section. Such a transducer was used to measure the output wave of the TM_{11} transducer and it was found that the TE_{11} component was down on the order of 30 db below the value which would be present if the radial field intensity observed at the top of Fig. 40 had been due to TE_{11} . By a process of elimination, therefore, and by virtue of the fact that we have a pure pattern suggesting the presence of a single mode, we have established that the mode generated is actually TM_{11} . Other checks can of course be made, such as measurement of the phase constant of the output wave.

The backward wave shown at the bottom of Fig. 40 has a maximum field more than 20 db below the maximum field of the forward wave and

has a six-peaked variation with angle which indicates the presence of TE_{31} .

The transfer loss of the TM_{11} transducer was derived by (1) calibrating the receiving probe on a known amount of power in the TE_{11} wave, (2) inserting this same amount of power in the rectangular waveguide of the coupled wave transducer and, with the probe at the transducer output, observing the change in the receiver response, and (3) correcting the observed loss using the theoretical difference in the radial electric field at the wall for the TE_{11} versus TM_{11} waves in the known waveguide diameter. (This technique is described in more detail by Aronoff.⁷) The result gave a transfer loss of about 25 db to the TM_{11} wave. The insertion loss for the rectangular guide of the transducer was less than 0.2 db.

Coupled-wave devices of the type shown in Fig. 39 were built for several of the modes in 2" round waveguide. The one built for the TE_{31} mode in 2" waveguide (mechanically similar to the TM_{11} model of Fig. 39) has several characteristics worthy of mention. Fig. 41 shows the

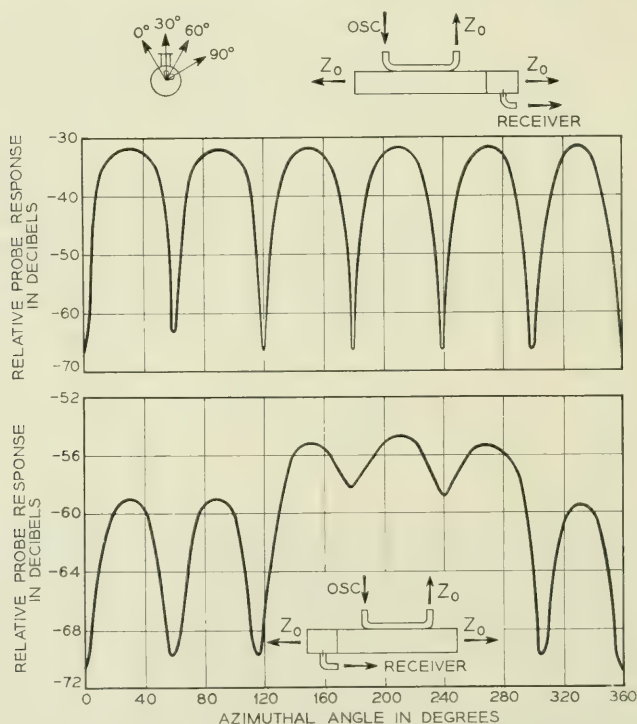


Fig. 41 — Distribution of radial electric field at the guide wall for the forward and backward waves of TE_{10} to TE_{31} coupled wave transducer.

TABLE II

Ratio of Forward Traveling TM_{11} Power to	Observed Discriminations		
	$\lambda_0 = 3.1$ cm	$\lambda_0 = 3.3$ cm	$\lambda_0 = 3.5$ cm
	db	db	db
TM_{11} Backward	>20	>20	>20
TE_{11} Forward	28.5	28	26
TE_{11} Backward	37	35	39
TM_{01} Forward	46	46	41.5
TM_{01} Backward	49	51	45.5
TE_{21} Forward	24.5	21	23
TE_{21} Backward	29.5	35	31
TE_{31} Forward	14	26	21.5
TE_{01} Forward	45	46	45
TE_{01} Backward	64	69	67

measured forward and backward wave patterns in the round guide, for excitation in one of the rectangular guides of the transducer. Only TE_{31} of the six modes possible in the 2" pipe at 3.3 cm has a six-lobed pattern of azimuthal distribution of radial electric field at the wall, and hence the clean pattern with equally spaced deep nulls indicates the presence of a rather pure TE_{31} mode. The six maxima of the forward wave were equal within ± 0.15 db. The backward wave had a peak electric field at least 23 db down on the peak electric field of the forward wave.

Using coupled transmission line techniques and the familiar geometric taper techniques, transducers were built for all of the six modes possible in 2" diameter pipe at 3.3 cm for use in the circular electric wave research program.¹ These transducers were used to measure the forward wave and backward wave output of the TM_{11} transducers, as given is Table II. In reality, imperfections in either one of the two transducers involved in a measurement could result in the recorded values of discrimination. For example, if the TM_{11} transducer were perfect and the TE_{01} output transducer contained some TM_{11} , then the insertion loss measurement involving the two transducers face to face would produce an indication of mode impurity. Since we do not have independent information on the mode purity of any one of the transducers at the level of the observed wave impurities, we can only state that both transducers involved in a discrimination measurement are probably at least as good as the number tabulated.

It should be noted that very high discriminations between TE_{01} and TM_{11} were achieved, despite the fact that this one discrimination depends solely on the mode-selective nature of the coupling orifice. Similar discriminations can be employed effectively to augment the wave-inter-

ference discrimination even in cases where there is difference between the desired and undesired modes' phase constants, to achieve very large discriminations. In the TM_{11} discriminations listed above, the values for TE_{31} are not great but are consistent with computed values for the coupling length and the coupling function employed; longer coupling lengths would produce better TM_{11} versus TE_{31} discriminations.

A TIGHTLY COUPLED $TE_{10}^{\square}$ TO TE_{01}° WAVE TRANSDUCER*

A highly efficient means of transferring power from dominant-mode rectangular waveguide to one of the higher modes of a multi-mode waveguide would be essential in a waveguide transmission system.¹ When several modes can propagate in one or both of the guides, the problem of achieving complete power transfer is more difficult and requires some new techniques. This section describes these techniques and gives experimental data for a circular-electric-wave ($TE_{10}^{\square} - TE_{01}^{\circ}$) transducer.

The desired transducer was required to make the wave transformation between a single-mode rectangular waveguide and the circular electric mode (TE_{01}°) of an 0.875" round waveguide at a nominal frequency of 24,000 mc. The 0.875" round waveguide at this frequency will support 10 modes of which the circular electric mode and its degenerate partner TM_{11}° are the fourth and fifth in order of appearance.

The minimum length of the coupling interval required to achieve mode discrimination may be estimated using loose coupling theory (equation 4). The mode nearest to TE_{01}° in phase constant is the TE_{31}° and for this mode a coupling length of about 0.18 meters is required in order to produce a value of θ/π equal to unity. As shown by equation (5) for uniform coupling, it is necessary to have θ/π equal to unity or greater in order to develop discrimination against the undesired mode.

The maximum coupling coefficient permissible for a given amount of mode impurity at the complete power transfer point may be estimated using the tight coupling theory of the preceding sections. For example, equations (31) and (32) show that for the ratio $(\beta_1 - \beta_2)/c$ equal to 10, the transfer loss to the undesired wave will always be greater than 14 db (regardless of the length of the coupling interval), corresponding to an energy loss for the desired wave of less than 0.2 db. For the TE_{01}° and TE_{31}° modes the calculated values of β_1 and β_2 lead to the conclusion that the coupling coefficient c between TE_{31}° and $TE_{10}^{\square}$ must be less

* When discussing the modes of hollow metallic waveguides of different cross-sectional shapes, it has been found convenient to use a superscript to designate the shape of the cross section. (See G. C. Southworth, *Principles and Applications of Waveguide Transmission*, D. Van Nostrand Co., 1950). Thus, $TE_{10}^{\square}$ refers to the TE_{10} mode in rectangular waveguide.

than 3.45 radians per meter. If the coupling coefficient for $TE_{10}^{\square}$ to TE_{01}° is equal to that for $TE_{10}^{\square}$ to TE_{31}° it follows that the total coupling length must be greater than 0.455 meters, because complete power transfer requires that the product of coupling-length times coupling-coefficient be exactly $\pi/2$ (see Fig. 17). Actually, the $TE_{10}^{\square} - TE_{31}^{\circ}$ coupling may be greater than the $TE_{10}^{\square} - TE_{01}^{\circ}$ coupling which leads to the requirement for longer coupling intervals. It is evident that the shorter coupling intervals may be employed at the sacrifice of greater mode impurities. The preceding calculations were made for the $TE_{10}^{\square} - TE_{31}^{\circ}$ and $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer ratios as though only one mode of the multi-mode waveguide were present at a time, i.e., using a theory based on coupling between two waves instead of a theory for the simultaneous coupling between a plurality of waves. It is felt that this is probably

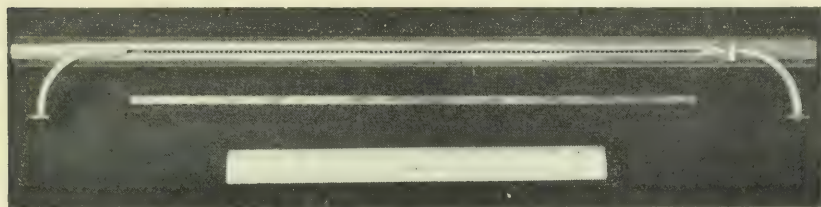


Fig. 42 — An experimental circular electric wave ($TE_{10}^{\square}$ to TE_{01}°) transducer for 24,000 mc.

justified provided that the coupling per unit length is weak and only one mode in each guide carries an appreciable amount of power.

Fig. 42 shows a photograph of one of the models used to obtain experimental data. The coupling holes were located in the narrow wall of the rectangular waveguide, thus avoiding coupling to all of the TM modes of the round waveguide. The total coupling length was 0.55 meters. The coupling orifices were spaced about 0.3 wavelengths in the dominant-mode rectangular waveguide, which assured reasonable directivity in the transfer of power between waveguides, provided that two or more coupling elements were employed.

The transfer loss between the rectangular waveguide and the circular electric mode of the round waveguide was measured as a function of the number of coupling elements, using the structure of Fig. 42 with the addition of a movable thin-walled metallic cylinder. The latter could be moved inside the transducer in such a way as to cover up a variable number of coupling holes, and contained a long wooden termination so that all the power entering the movable cylinder was absorbed. The inner diameter of the movable cylinder was large enough to propagate the

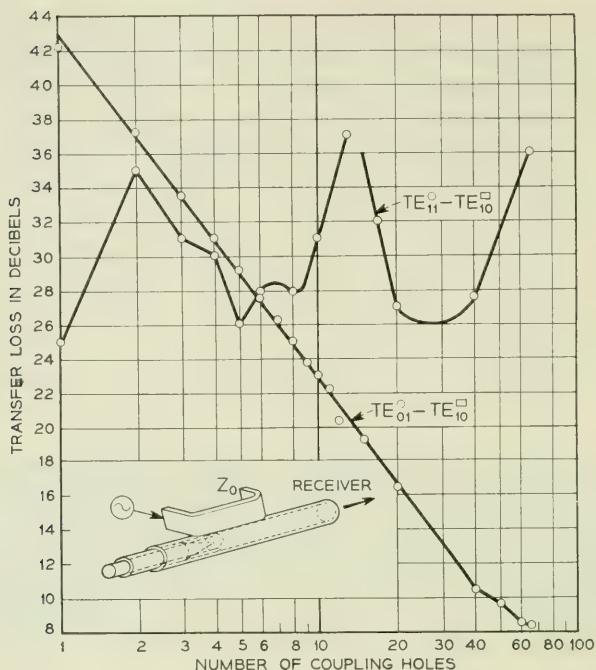


Fig. 43 — Transfer losses versus frequency for the transducer of Fig. 42.

circular electric wave but did cut off some of the waves which could propagate in the round guide of the transducer itself. The measured transfer loss under these conditions is recorded in Fig. 43. It is seen that the $TE_{10}^{\square} - TE_{01}^{\circ}$ coupling was so weak as to be in the region where power from successive coupling elements should add inphase all the way up to 40 coupling elements. The observations show the inphase addition for less than 30 coupling elements but show a marked deviation in the vicinity of 40 to 66 coupling elements. This is evidence of inequality of the phase constants for the TE_{01}° and $TE_{10}^{\square}$ waves. More will be said about this matter presently. The transfer loss between the rectangular waveguide and the TE_{11} mode of round waveguide, is also recorded in Fig. 43. As expected, the power from successive coupling elements did not add inphase and no appreciable build-up of power in the TE_{11} mode took place.

One way of evaluating the total power in all modes other than the circular electric mode, is to measure the value of the transverse magnetic intensity at the wall of the round waveguide. The circular electric wave has no such field component and all other waves do possess such a field

component. Thus the total value of the transverse magnetic intensity at the round waveguide wall is a measure of the impurity associated with the circular electric wave. (This is very similar to the radial probe technique described by M. Aronoff.⁷) Using this method of evaluation, the mode impurities present at the output of the transducer were measured as a function of the number of coupling elements, and the results are recorded in Fig. 44. The absolute calibration of the ordinate relates the observed magnetic intensity to that which the same power input used at the rectangular guide would have produced if placed in the round waveguide in the TE_{11} mode. These measurements show that for all of the modes other than the circular electric mode, the energy components from successive coupling elements suffer destructive interference. Although curves are shown only for one and for 66 coupling elements, the patterns for intervening numbers of coupling elements were similar in shape and never exceeded an intensity value greater than about 6 db above that given for the 66 coupling element case; thus the mode discriminating property of the coupled wave transducer was verified experimentally.

Returning to the question of $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer loss, it is clear from Fig. 43 that the rectangular waveguide has a phase constant which is not equal to that of the circular electric mode in the round waveguide. One reason for this inequality lies in the fact that the coupling elements disturb the phase constant in the two waveguides unequally, a consequence of the fact that some of the power transferred to the round wave-

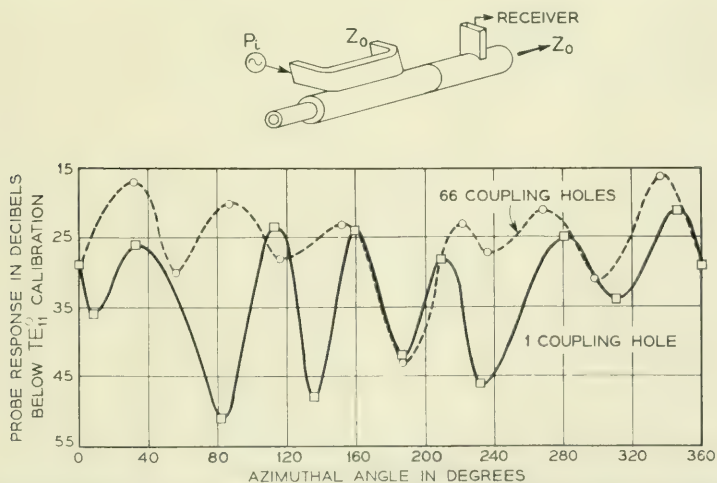


Fig. 44 — Distribution of transverse magnetic intensity at the wall for the transducer of Fig. 42.

guide on a single coupling element basis, appears in modes other than TE_{01} . Thus, the total coupling to $TE_{10}^{\square}$ is greater than to TE_{01}° . The total coupling modifies the phase constant of each line, per (20'), and since the total coupling coefficient is unequal for the $TE_{10}^{\square}$ and the TE_{01}° modes, the perturbed phase constants should be expected to be unequal when the unperturbed phase constants are made equal. A method of determining the magnitude of this phase-constant disturbance has been suggested by S. A. Schelkunoff. In this method the reflected wave from a single coupling orifice is measured in the dominant waveguide and in the single mode of interest in the multi-mode waveguide. Having defined the ratio of the incident to the reflected power in the same mode by the symbol p , Schelkunoff determines that the disturbed phase constant β' , is related to the undisturbed phase constant β by the relation

$$\beta' = \beta + \sqrt{\frac{p}{d}}, \quad (41)$$

in which " d " is the distance between the coupling orifices in the coupling arrangement which one wishes to evaluate. This relation may be used to evaluate the change in the phase constant for the circular electric mode and for the wave in the dominant waveguide, and the change of wave-

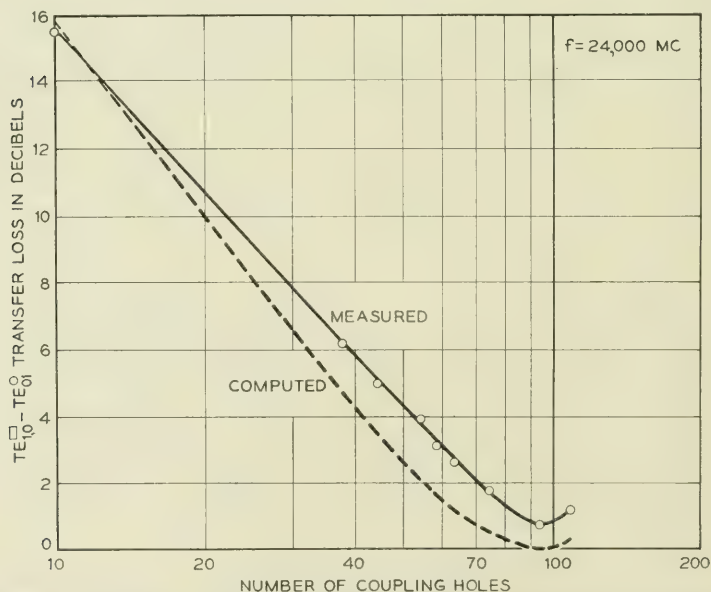


Fig. 45 — Transfer loss for the transducer of Fig. 42 after increasing the coupling hole sizes and correcting the phase constant.

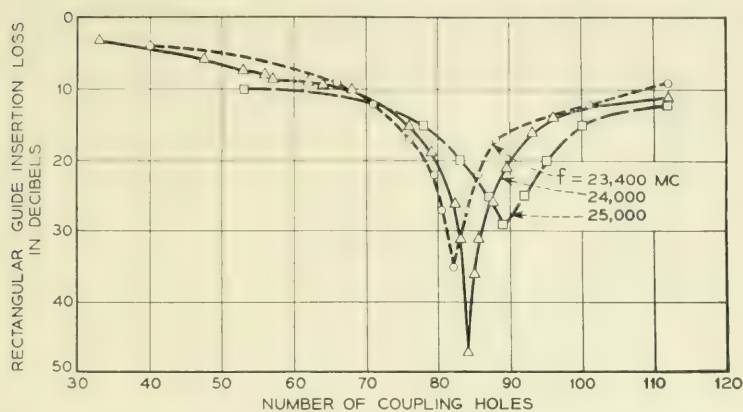


Fig. 46 — Rectangular guide insertion loss for the transducer of Fig. 42.

guide dimensions required to correct this phase constant difference may be computed as though the coupling elements were not present.

For the small phase constant disturbances which are associated with the weak couplings employed, this procedure was found very accurate. The reflection measurements and associated calculations for the model of Fig. 42 indicated that the rectangular guide width should be 0.340" for equality of phase constants instead of 0.359" as computed neglecting coupling effects. The measured value of the transfer loss when the individual coupling holes had been enlarged and the rectangular guide width had been altered to the 0.340" value is shown in Fig. 45. It is evident that the theoretical value of 0 db transfer loss was approached, and that the shape of the transfer loss versus number of coupling elements, was reproduced very well. The 0.75 db minimum transfer loss consisted of no more than 0.3 db heat loss, the remaining loss being due to power present in other modes.

The measured insertion loss in the rectangular waveguide is shown as a function of the number of coupling holes at the three frequencies in Fig. 46. Complete power transfer would, of course, correspond to an infinite insertion loss in the rectangular waveguide. It is interesting to note that at 24,000 mc the peak in the rectangular guide insertion loss occurred at 85 coupling elements whereas the maximum in the $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer loss characteristic occurred at about 96 coupling elements (Fig. 45). This difference is likely to be the result of power transferred back to the rectangular waveguide from round waveguide modes other than circular electric. Additional evidence of deviations due to the coupling between a plurality of waves was obtained; the rectangular-guide insertion loss as a function of number of coupling elements did not increase

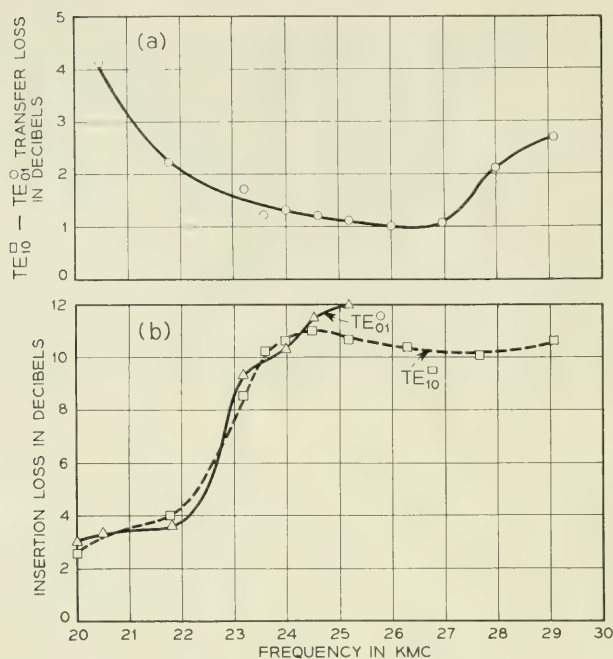


Fig. 47 — Transfer and insertion loss versus frequency in the utilized modes of the transducer of Fig. 42 using all (112) coupling holes.

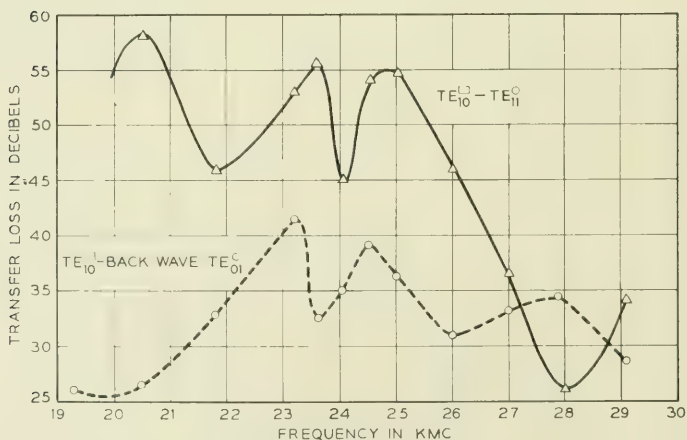


Fig. 48 — Circular electric wave directivity and one unwanted mode (TE_{11}°) output versus frequency for the transducer of Fig. 42.

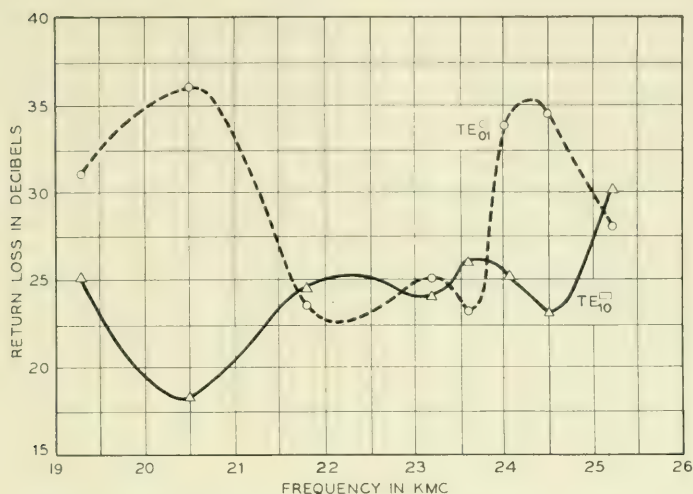


Fig. 49 — Impedance characteristic of the transducer of Fig. 42.

smoothly according to a cosine amplitude function as would be expected for two coupled waves of identical phase constant, but instead exhibited ripples. The remarkable thing about the data of Figs. 45 and 46 is that it agrees with the theory for two coupled waves as well as it does.

The coupling per individual orifice decreases with increasing frequency and this is verified by the observation (Fig. 46) that a greater number of coupling elements are required to reach the maximum insertion loss in the rectangular guide at the higher frequency.

Some indication of the overall bandwidth of this first experimental model is given in Figs. 47, 48 and 49 which show respectively the $TE_{10}^{\square} - TE_{01}^{\circ}$ transfer loss, the insertion losses in the $TE_{10}^{\square}$ and TE_{01}° modes, the $TE_{10}^{\square} - TE_{11}^{\circ}$ and $TE_{10}^{\square} -$ backward wave TE_{01}° transfer losses, and the $TE_{10}^{\square}$ and TE_{01}° return losses in the frequency range 20,000 to 30,000 mc. No one of these characteristics represents the degree of excellence which is achievable but they do demonstrate that good impedance match, low transfer losses to the desired mode, and appreciable discrimination against unwanted modes, can be achieved over frequency ratios on the order of 1.5.

FREQUENCY SELECTIVITY

In the case wherein the coupling is so weak as to not affect the total phase constant appreciably, all modes of hollow conductor waveguides of any cross section have the same phase constant at all frequencies provided that these modes have the same cut-off frequency. This results

in very broad band mode-selective characteristics, as has been demonstrated.

The transfer loss characteristics are in general a function of frequency, since the individual coupling holes are somewhat frequency selective. There may be applications wherein less variation in transfer loss as a function of frequency is required. One approach to this problem is to make the coupling holes individually have less coupling variation with frequency; since the total coupling loss between two identical transmission lines is a function only of number of coupling holes and the loss per hole (equations (39) and (40)) constant coupling per hole will produce constant coupling overall. Riblet and Saad⁶ have reported on this approach.

There is another approach to obtaining flat coupling versus frequency despite variations in the coupling per hole, and that is to intentionally create a difference between the phase constants of the two coupled lines. Fig. 17 illustrates the transfer characteristic when the coupled lines have unequal phase constant, and either identical or negligible attenuation constants. Near the maximum for the transferred wave $|E_2^*|$ there is a region wherein the transfer loss is independent of coupling strength, and the transfer loss in this flat-loss region is under control of the ratio $(\beta_1 - \beta_2)/c$. Hence for a given transfer loss there is an optimum ratio of phase constant difference to coupling strength in order to minimize the overall transfer loss variation. For the distributed coupling case, equations (31) and (32) represent the transferred wave amplitude and show that the transferred wave goes through a maximum as a function of integrated coupling strength cx , when

$$\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1} cx = \frac{\pi}{2} + n\pi. \quad (42)$$

The transferred amplitude at this maximum point is

$$E_{2\max}^* = \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}}. \quad (43)$$

The integrated coupling strength at the maximum point is

$$c_0x_0 = \frac{1}{\sqrt{\frac{(\beta_1 - \beta_2)^2}{4c^2} + 1}} \cdot \frac{\pi}{2}. \quad (44)$$

For the important case of an optimum 3 db transfer loss coupler, E_2^* is 0.707. Then $(\beta_1 - \beta_2)/c$ equals 2 and c_0x_0 equals $\pi/2\sqrt{2}$ from (43)

and (44). Assuming a coupling length x_0 of two wavelengths in the line with the smaller phase constant, it follows that β_1/β_2 is about 1.18 showing that a phase-constant difference of 18% is required. This phase-constant difference is quite readily attainable in the waveguide structure of Fig. 50(a). The two modes coupled together are given slightly different cut-off wavelengths in the coupling region, and may be tapered to the standard waveguide size outside the coupling region. The desired phase-constant difference can also be obtained in two identical metallic guides in the coupling region as sketched in Fig. 50(b). Although rectangular waveguides are used in Fig. 50 to illustrate the method of obtaining frequency independent transfer characteristics, the approach is general and may be applied to any form of single or multi-mode transmission line.

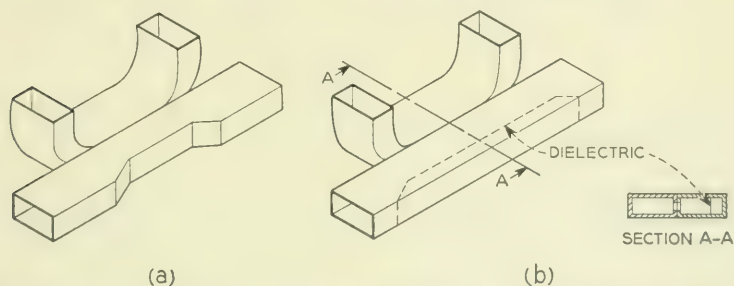


Fig. 50 — Examples of structures in which flat transfer loss may be obtained despite coupling loss variations.

In either dominant-mode directional couplers or in multi-mode coupled-wave devices such as the one illustrated in Fig. 1, one may obtain much more frequency selectivity than occurs incidentally due to the frequency sensitivity of the coupling elements used. This may be done by coupling two transmission lines which have the same phase constant at one frequency, but unequal phase constants at other frequencies. Then, as shown by equation (31), the midband transfer loss may be set at any desired value by adjusting the integrated coupling strength $c\bar{x}$ at midband (where $\beta_1 - \beta_2 = 0$), and at other frequencies where $(\beta_1 - \beta_2) \neq 0$, the transfer loss will increase. For the particular case of $c\bar{x} = \pi/2$ (fixed) for which complete power transfer occurs when $\beta_1 = \beta_2$ (and assuming $\alpha_1 = \alpha_2$ or both α 's are negligible), Fig. 51 shows the shape of the filter characteristic, E_2^* versus $(\beta_1 - \beta_2)/2c$. This plot is valid for any form of transmission line.

A very simple configuration for realizing such a frequency-selective filter involves coupling between two hollow conductor waveguides, one

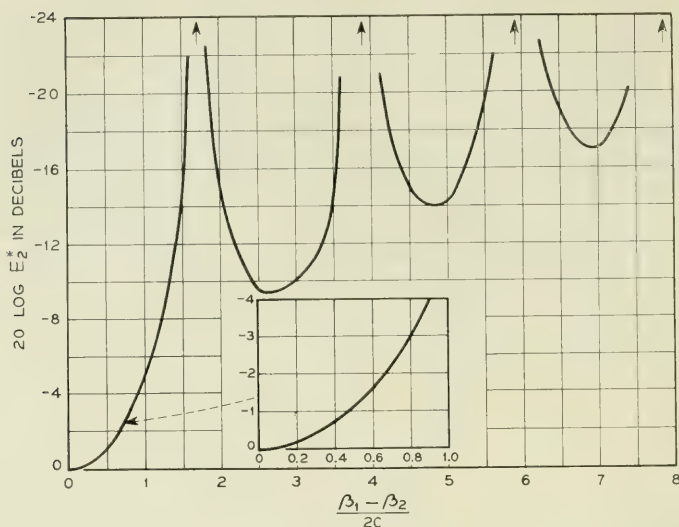


Fig. 51 — Transfer loss E_2^* versus $(\beta_1 - \beta_2)/2c$ for coupling strength $cx = \pi/2$, the value required for complete power transfer.

of which is air-filled and the other of which is filled with a material of dielectric-constant ϵ . The phase constants for these waveguides have the form sketched in Fig. 52, in which β_0 is the phase constant in free space. At the frequency f_m the two waveguides have identical phase constants and, in a typical case, negligible loss constants so that complete power transfer can be obtained. For the case $\epsilon = 2.55$, Fig. 53 shows the computed frequency characteristic on the assumption that the integrated coupling is set for complete transfer ($cx = \pi/2$) and is independent of frequency. (Actually the usual coupling mechanisms are somewhat frequency sensitive and would increase the selectivity somewhat.) This filter

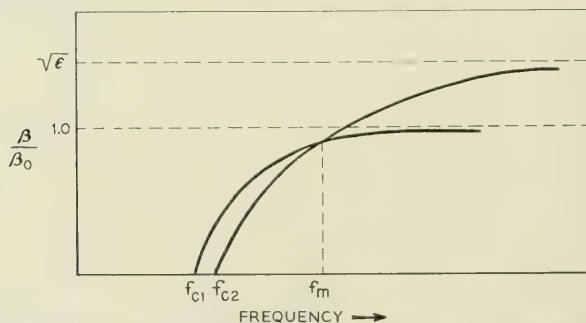


Fig. 52 — The general form of the phase constants for two hollow conductor waveguides, one of which is filled with a dielectric.

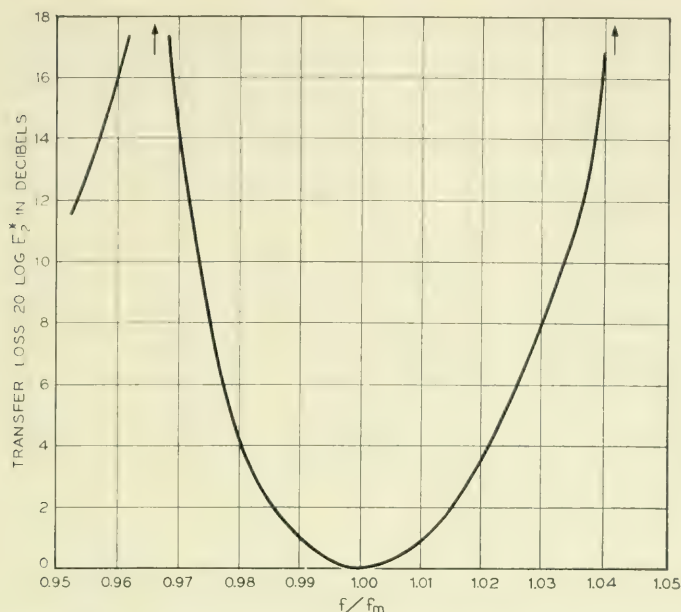


Fig. 53 — The transfer loss E_2^* versus normalized frequency for two coupled hollow conductor waveguides, one of which is air filled and has a guide wavelength $\sqrt{2}$ times the free space wavelength at f_m , and the other of which is filled with a material of dielectric constant 2.55 with dimensions chosen for equality of phase constant with the air-filled guide at f_m . Coupling cx assumed constant at $\pi/2$.

characteristic applies regardless of the shapes of the hollow conductor waveguides (which may be dissimilar) and regardless of the modes selected.

It is apparent that frequency selectivity in the transfer characteristic E_2^* can also be obtained without requiring that the phase constants be unequal by using coupling elements which are frequency sensitive.

DIELECTRIC WAVEGUIDE CONFIGURATIONS

The coupled-wave approach to circuit design is applicable using any form of transmission line, the only important variant associated with different forms of line being the physical structure associated with introducing the desired coupling between lines. In a recent publication⁸ A. G. Fox showed that dielectric waveguides are very attractive for use in the millimeter wavelength range, and this section points out how dielectric waveguides can be used in various forms of coupled wave devices. Fox showed that dielectric waveguides arranged in the configuration sketched in Fig. 54 are coupled by the electric field components only, and that

periodic energy exchange of the type described by equations (26) and (27) is observed. Moreover, he also showed that if one line were made very lossy the energy exchange phenomena disappeared and, despite sufficient coupling to cause complete power transfer when both lines were loss-free, power passed through the coupling region in the low-loss line with less than 0.25 db attenuation. This verified the predictions of equations (35) and (36).

Other implications of the coupled wave theory can also be utilized in dielectric waveguides. If the two lines (Fig. 54) are made of materials having different dielectric constants and their cross-sectional dimensions set so as to secure identical phase constants at a frequency f_m , then a frequency-selective coupled-wave filter results and the selectivity characteristic of Fig. 53 applies. As an alternative to using materials having different dielectric constants, the same dielectric may be used for both lines by making one line solid and the other hollow.

If both lines are made of the same material and the cross-sectional dimensions are set so as to obtain a known difference between their phase constants, the result is a directional coupler having a region of flat transfer loss (of any desired magnitude) and equations (42), (43) and (44) apply.

Both of the preceding applications can be carried out in dielectric waveguides having arbitrary cross-sectional shapes.



Fig. 54 — Coupled dielectric waveguides.

If one of the transmission lines (Fig. 54) is round and the other is rectangular and if their cross-sectional dimensions are set for equal phase constants, then the power in one of the two polarizations of the round line may be transferred to any desired extent to the rectangular guide, and power in the other polarization of the round guide will pass the coupling region undisturbed. Two such rectangular-rod to round-rod coupling configurations arranged in cascade along the round-rod, with the two rectangular rods coupled in planes at 90° to each other, constitutes a means for independently connecting to the two polarizations of the round-rod. This type of device depends upon the fact that the phase constants of the two polarizations of round-rod are identical, whereas the two phase constants for the rectangular rod are different. Thus a wave interference occurs in the transfer characteristic for one of the polarizations, and for suitable values of $(\beta_1 - \beta_2)/c$ (see Fig. 18) the power transferred in this polarization can be made small.

SUMMARY

Two approaches to a theoretical description of the behavior of two coupled waves have been presented. One, based on the assumption of negligibly small coupling, is applicable in cases where very little power is transferred between the coupled waves. The other, a solution based on uniform coupling between waves in the coordinate of propagation, is valid for any magnitude of total coupling.

The loose coupling theory shows how to taper the coupling distribution in order to minimize the length of the coupling interval required for a given degree of directivity and/or for a given magnitude of mode impurity. In particular, it is possible to shape the coupling distribution so as to discriminate sharply against one or more undesired modes in a coupled-wave arrangement involving just a few modes. (See Figs. 7 and 15 for examples).

The theory indicates that significant exchange of power takes place provided that the attenuation and phase constants of the coupled waves are equal, or provided that the difference between the attenuation constants and the difference between the phase constants are small compared to the coefficient of coupling. A suitable difference between either the attenuation constants or the phase constants of two coupled waves is sufficient to prevent appreciable energy exchange (equations 29-32 and 35-36).

It follows that substantially single-mode propagation is possible in a multi-mode structure even though geometrical effects tending to cause coupling between modes are present. A gradual transition in the boundary

of a multi-mode waveguide will not cause an appreciable exchange of power between modes provided that the quantity $(\beta_1 - \beta_2)/c$ is sufficiently large for the modes which are coupled by the boundary change. Similarly, for disturbances in the coupled-wave system which takes place over a large number of wavelengths in the direction of propagation, the coupled-wave theory indicates that all conversion will take place in the forward direction and very little reflection in any mode will result.

The tight coupling theory shows that for the case of identical complex propagation constants, a periodic exchange of energy between waves takes place along the coordinate of propagation. The only effect of the existence of an attenuation constant for both waves (compared to the dissipationless case) is to add the same exponential attenuation factor (to the periodic energy exchange phenomenon) which would have existed for a wave traveling on one of the lines in the uncoupled state.

When the phase constants of the two coupled waves are not equal (and the attenuation constants are either equal or negligibly small compared to the coupling coefficient), the exchange of energy between waves is no longer complete but remains periodic (Fig. 17). The quantity $(\beta_1 - \beta_2)/c$ determines the fraction of the total energy which is exchanged, and also modifies the period of the energy exchange phenomenon along the axis of propagation.

When the phase constants of the two lines are equal but the attenuation constants are unequal, the energy transfer phenomenon differs only slightly from that associated with equal propagation constants provided that the quantity $(\alpha_1 - \alpha_2)/c$ is less than about -0.1 . For $(\alpha_1 - \alpha_2)/c$ more negative than about -1 , the periodicity of the energy transfer phenomenon has largely disappeared (Fig. 23) and as $(\alpha_1 - \alpha_2)/c$ becomes on the order of -10 or more, the principal effect of the coupling for the low loss line is a minor alteration of the phase and attenuation constants. The wave amplitude for unit input on the low-loss line becomes [from (33) for $|(\alpha_1 - \alpha_2)/c| \gg 1$]

$$E_1 = \epsilon^{-[\alpha_1 - c^2/(\alpha_1 - \alpha_2) + i(c + \beta)]x} \quad (45)$$

Through proper choice of the phase constants relative to the coupling coefficient in two coupled transmission lines, it is possible to make directional couplers having an arbitrary transfer loss that is independent of frequency despite variations in coupling strength with frequency (equations 43-44). It was also shown that the coupled-wave approach may be utilized to create highly frequency-selective filters which may operate between single-mode media or between selected individual modes of a multi-mode system.

The experimental data given for two dominant-mode rectangular waveguides showed that the periodic energy exchange theoretically predicted for a coupled-wave system can be achieved in coupled transmission lines.

Performance characteristics were given for some loosely coupled transducers between a dominant-mode rectangular waveguide and one mode of a six-mode waveguide. A tapered coupling distribution was used to achieve the mode selectivity in a limited length interval.

The problems associated with a coupled-wave transducer for transferring all of the power from a dominant-mode rectangular waveguide to the circular electric mode in a ten mode waveguide, were discussed and the observed characteristics of an experimental model were given.

The application of coupled-wave techniques to other types of transmission systems was illustrated by pointing out analogous structures using coupled dielectric waveguides.

ACKNOWLEDGMENT

The writer is indebted to W. W. Mumford for helpful discussions during the early stages of this work; to R. W. Dawson who made most of the measurements on the models of Figs. 36 through 44, and to G. D. Mandeville who made measurements on the model of Fig. 39.

BIBLIOGRAPHY

1. S. E. Miller and A. C. Beck, Low-Loss Waveguide Transmission, *Proc. I.R.E.*, **41**, pp. 348-358, Mar., 1953.
2. S. E. Miller, Notes on Methods of Transmitting the Circular Electric Wave Around Bends, *Proc. I.R.E.*, **40**, pp. 1104-1113, Sept., 1952.
3. W. W. Mumford, Directional Couplers, *Proc. I.R.E.*, **35**, pp. 160-165, Feb., 1947.
4. C. L. Dolph, A Current Distribution for Broadside Arrays Which Optimizes the Relation Between Beam Width and Side Lobe Level, *Proc. I.R.E.*, **34**, pp. 335-348, June 1946.
5. S. E. Miller and W. W. Mumford, Multi-Element Directional Couplers, *Proc. I.R.E.*, **40**, pp. 1071-1078, Sept., 1952.
6. H. J. Riblet and T. S. Saad, A New Type of Waveguide Directional Coupler, *Proc. I.R.E.*, **36**, pp. 61-64, Jan., 1948.
7. M. Arnoff, Radial Probe Measurements of Mode Conversion in Large Round Waveguide with TE_{01} Excitation, (submitted to *Proc. I.R.E.*).
8. A. G. Fox, New Guided Wave Techniques for the Millimeter Wavelength Range, given orally at the March, 1952, I.R.E. National Convention. To be submitted to the Proceedings.
9. Alan A. Meyerhoff, Interaction Between Surface-Wave Transmission Lines, *Proc. I.R.E.*, **40**, pp. 1061-1064, Sept., 1952.
10. P. E. Krasnushkin and R. V. Khokhlov, Spatial Beating in Coupled Waveguides *Zh. Tekh. Fiz.*, **19**, pp. 931-942, Aug., 1949, (in Russian).
11. W. J. Albersheim, Propagation of TE_{01} Waves in Curved Wave Guides, *B.S.T.J.*, **27**, pp. 1-32, Jan., 1949.

Theoretical Fundamentals of Pulse Transmission — I

By E. D. SUNDE

(Manuscript received September 23, 1953)

A compendium is presented of theoretical fundamentals relating to pulse transmission, for engineering applications. Emphasis is given to the consideration of various imperfections in transmission systems and resultant transmission impairments or limitations on transmission capacity.

In Part I of this paper, Sections 1 to 11, fundamental properties of transmission-frequency characteristics are discussed, together with general relations between frequency and pulse transmission characteristics and special transmission characteristics of importance in pulse systems. This is followed by a presentation of engineering methods of evaluating pulse distortion from various types of gain and phase deviations.

In Part II, Sections 12–16, transmission limitations imposed by characteristic distortion will be discussed.

PART I

1. Properties of Transmission-Frequency Characteristics.....	724
2. Frequency and Pulse Transmission Characteristics.....	730
3. Idealized Characteristics with Sharp Cutoff.....	736
4. Idealized Characteristics with Gradual Cutoff.....	741
5. Idealized Characteristics with Natural Linear Phase Shift.....	743
6. Pulse Echoes from Phase Distortion.....	752
7. Pulse Echoes from Amplitude Distortion.....	761
8. Fine Structure Imperfections in Transmission Characteristics.....	764
9. Transmission Distortion by Low-frequency Cutoff.....	773
10. Transmission Distortion by Band-edge Phase Deviations.....	779
11. Band-pass Characteristics with Linear Delay Distortion.....	784

PART II

12. Impulse Characteristics and Pulse Train Envelopes	
13. Transmission Limitations in Symmetrical Systems	
14. Transmission Limitations in Asymmetrical Sideband Systems	
15. Double vs. Vestigial Sideband Systems	
16. Limitation on Channel Capacity by Characteristic Distortion	
Acknowledgements	
References	

INTRODUCTION

Pulse transmission is a basic concept in communication theory and certain methods of modulating pulses to carry information approach in their characteristics the ideal performance allowed by nature. In certain applications, such as telegraphy, pulse signalling and data transmission, it has the advantage of great accuracy, since the information is transmitted in digital form by "on-off" pulses. This at the same time facilitates regeneration of pulses to avoid accumulation of distortion from noise and other system imperfections, together with the storing, automatic checking and ciphering of messages, as well as their translation into different digital systems or transmission at different speeds, as may be required in extensive communication systems. Another characteristic of pulse systems is that improved signal-to-noise ratio can be secured in exchange for increased bandwidth, as in pulse code, pulse position and certain other methods of pulse modulation. Finally, pulse modulation systems permit multiplexing of communication channels on a time division basis, which under appropriate conditions may have appreciable advantages over frequency division in the design of multiplex terminals.

In pulse modulation systems, pulses are applied at the transmitting end in various combinations, or in varying amplitude, duration or position, depending on the type of system. Pulses thus modulated to carry information may be transmitted in various ways, or undergo a second modulation process suitable to the transmission medium. The received pulses will differ in shape from the transmitted pulses because of bandwidth limitations, noise and other system imperfections. The performance of the system in the absence of noise can be predicted if the "pulse transmission characteristic" is known, that is, the shape of a received pulse for a given applied pulse.

Although the pulse-transmission characteristic suffices for determination of system performance it is customary for various reasons to relate it to the "transmission-frequency characteristic," that is, the steady-state transmission response expressed as a function of frequency. For one thing the transmission-frequency characteristics of various existing facilities and their components are known, and for new facilities can be determined more readily by calculation or measurements than the pulse-transmission characteristic. But the more fundamental reason is that the transmission-frequency characteristics of various system components connected in tandem or parallel can readily be combined to obtain the over-all transmission characteristic, while this is not the case for pulse transmission characteristics. It is thus possible to analyze complicated systems with the transmission-frequency characteristic as a basic

parameter, and to specify requirements that must be imposed on the transmission-frequency characteristic of the system and its components for a given transmission performance.

A fundamental problem in pulse modulation systems is transmission distortion of pulses by system imperfections in the form of phase and gain deviations over the transmission band or a low-frequency cut-off, usually referred to as "characteristic distortion," which may give rise to excessive interference between pulses and resultant crosstalk noise or errors in reception, depending on the type of system. Because of such interference, characteristic distortion limits the number of pulse amplitudes permissible in the transmission of information or messages over a given channel, and may reduce the rate at which pulses can be transmitted in systems employing only two pulse amplitudes, the minimum number. It thus places a limitation on channel capacity which, unlike signal distortion by noise, cannot be overcome by increasing the signal power.

Characteristic distortion is an important consideration particularly in wire systems where there is a low-frequency cut-off caused by transformers, and where the transmission band may extend over several octaves with substantial variation in attenuation and phase shift, or may be sharply confined by filters. In wire systems there are also fine structure deviations from a smooth attenuation and phase characteristic of a more or less random nature, resulting from small random impedance variations and mismatches along the lines. Gain and phase deviations remaining even after fairly elaborate equalization may be appreciable and difficult to overcome, especially in systems comprising a large number of repeater sections.

The purpose of this paper is to present a compendium of theoretical fundamentals on pulse transmission in a form suitable for engineering applications, both from the standpoint of design of new pulse transmission systems and pulse transmission over existing facilities. Emphasis is placed on considerations of various system imperfections, because of their importance from the standpoint of transmission performance, and since literature on this question is rather limited. Certain fundamental properties of transmission-frequency characteristics are discussed, together with general relations between frequency and pulse transmission characteristics and special transmission characteristics of importance in pulse systems. This is followed by a presentation of methods of evaluating pulse distortion from various types of gain and phase deviations, together with resultant transmission impairments or limitations on pulse transmission rates in low-pass, symmetrical and asymmetrical sideband

systems. Conversely, these methods may be used in the design of pulse modulation systems to evaluate requirements imposed on the transmission characteristics for a given transmission performance.

Transmission impairments may result from system imperfections other than characteristic distortion, which require a different theoretical approach and are not considered here. Among them are erratic timing of pulses, thermal and other noise within the transmission system and interference from outside sources, such as other communication systems or atmospheric disturbances.

1. PROPERTIES OF TRANSMISSION-FREQUENCY CHARACTERISTICS

A basic parameter of transmission systems is the transmission-frequency characteristic

$$T(i\omega) = A(\omega)e^{-i\psi(\omega)}, \quad (1.01)$$

in which $\omega = 2\pi f$ is the radian frequency, $A(\omega)$ is the amplitude and $\psi(\omega)$ the phase characteristic. The transmission-frequency characteristic may designate the ratio of received voltage to transmitted current, of received current to transmitted voltage, of received to transmitted current or of received to transmitted voltage. The two latter ratios are not the same except for symmetrical networks with impedance matching at both ends. For symmetrical structures having appreciable attenuation, such as transmission lines between repeaters, the ratios are virtually the same with impedance matching at the receiving end. In the following, $T(i\omega)$ will designate any of the above ratios, as the case may be.

When a number of networks are connected in series, as is usually the case in transmission systems, the resultant transmission characteristic is

$$\begin{aligned} T(i\omega) &= T_1(i\omega) T_2(i\omega) \cdots T_n(i\omega), \\ &= (A_1 A_2 \cdots A_n) e^{-i(\psi_1 + \psi_2 + \cdots + \psi_n)}, \end{aligned} \quad (1.02)$$

where $T_1, T_2, \cdots, T_n$ are the transmission characteristics of the individual networks with the same impedance terminations as encountered in the series arrangement, i.e. as measured in place or with equivalent terminations.

The phase characteristic ψ can in general be regarded as the sum of three components. The first is the minimum phase shift component, ψ^0 , which has a definite relation to the amplitude characteristic of the system, and is of particular interest in connection with phase distortion with different types of amplitude characteristics. The second is a

linear component $\omega\tau_d$, which represents a constant transmission delay τ_d for all frequencies, as in the case of an ideal delay network. Ladder type structures and transmission lines have phase characteristics which can be represented by the above two components. The third component can be represented by a lattice structure with constant amplitude characteristic but varying phase. Such a network component may be present in a transmission system or may be inserted intentionally for phase equalization, i.e. to supplement the first component above so as to secure a linear phase characteristic without altering the amplitude characteristic of the system.

The following discussion is concerned with the relationship of the first component to the amplitude characteristic of the system, or conversely.

The natural logarithm of the transmission-frequency characteristic given by (1.01) is

$$\ell n T(i\omega) = \ell n A(\omega) - i\psi(\omega). \quad (1.03)$$

The component $\ell n A(\omega)$ is referred to as the attenuation characteristic, and when expressed in decibels equals $8.69 \ell n A(\omega)$.

The following relations exist between the attenuation and phase characteristics of minimum phase shift systems or system components:^{1,2}

$$\ell n A(\omega) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\psi^0(u)}{\omega - u} du = \frac{2}{\pi} \int_0^{\infty} \frac{u\psi^0(u)}{u^2 - \omega^2} du, \quad (1.04)$$

and

$$\psi^0(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\ell n A(u)}{\omega - u} du = -\frac{2}{\pi} \int_0^{\infty} \frac{\omega \ell n A(u)}{u^2 - \omega^2} du. \quad (1.05)$$

In the evaluation of these integrals, the principal values are to be used, i.e., results of the form $\ell n(-u)$ are to be taken as $\ell n|-u|$ rather than $\ell n|u| + i\pi$.

As an example consider an attenuation characteristic as shown in Fig. 1, with $A(\omega) = A_0$ between $\omega = 0$ and ω_c and A_1 between $\omega = \omega_c$ and ∞ . Equation (1.05) then becomes

$$\begin{aligned} \psi^0(\omega) &= -\frac{2\omega}{\pi} \left[\ell n A_0 \int_0^{\omega_c} \frac{du}{u^2 - \omega^2} + \ell n A_1 \int_{\omega_c}^{\infty} \frac{du}{u^2 - \omega^2} \right], \\ &= \frac{1}{\pi} \ell n(A_0/A_1) \ell n \left| \frac{\omega_c + \omega}{\omega_c - \omega} \right|. \end{aligned} \quad (1.06)$$

In Fig. 1 is shown the phase characteristic for $A_0/A_1 = 100$, corresponding to a 40 db cutoff at $\omega = \omega_c$.

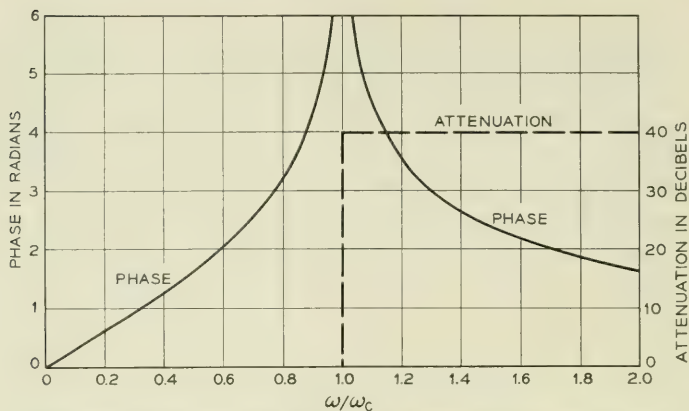


Fig. 1 — Low-pass transmission frequency characteristic with sharp cut-off.

In Fig. 2 the attenuation and phase characteristics are shown as a function of ω/ω_c for $\omega < \omega_c$ and as a function of the inverse ratio ω_c/ω for $\omega > \omega_c$. It will be noticed that for the above case the phase characteristic is infinite for $\omega/\omega_c = 1$ and has even symmetry about this point, while the attenuation characteristic has odd symmetry with respect to the midpoint of the amplitude discontinuity. The phase characteristic may be modified by a gradual cutoff in the attenuation characteristic, as illustrated in the figure. It is possible to shape the attenuation characteristic to obtain a linear phase characteristic in the transmission band, i.e. between $\omega/\omega_c = 0$ and 1. Since transmission systems with a

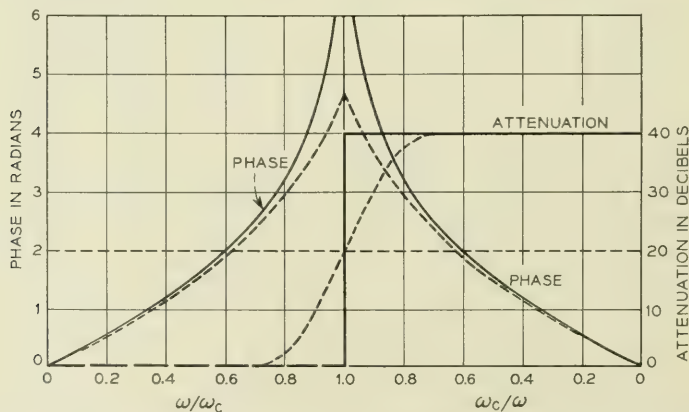


Fig. 2 — Solid curves same as in Fig. 1, but with inverse scale for $\omega/\omega_c > 1$. Dashed curves illustrate modification in phase characteristic with gradual cut-off in attenuation (not computed).

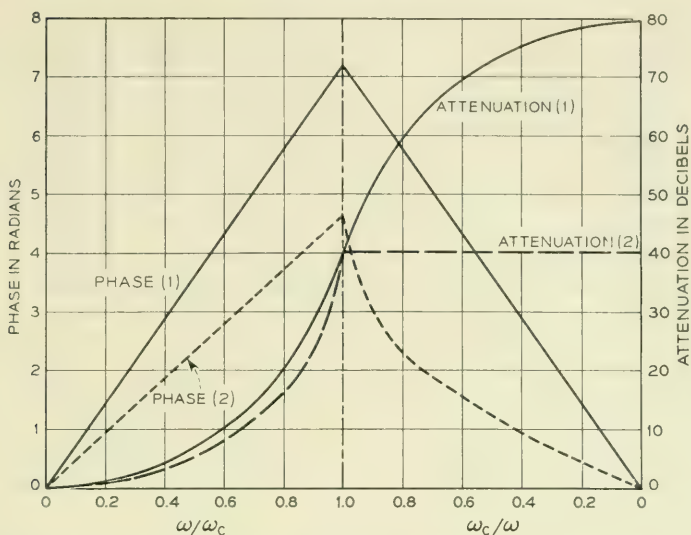


Fig. 3 — Low-pass transmission frequency characteristics with natural linear phase shift for $\omega/\omega_c < 1$.

linear phase characteristic in this range are of particular importance in pulse transmission, this case will be considered further.

It will be assumed that the phase characteristic has even symmetry when expressed in the scales of Fig. 2, in which case the phase characteristic as shown by the solid lines in Fig. 3 is given by

$$\begin{aligned}\psi^0(\omega) &= \omega\tau & \omega/\omega_c < 1, \\ &= \omega_c^2\tau/\omega & \omega/\omega_c > 1.\end{aligned}\quad (1.07)$$

With these expressions in (1.04) the attenuation characteristic becomes:

$$\ln A(\omega) = \frac{2\omega_c\tau}{\pi} \left[1 + \frac{1}{2} \left(\frac{\omega_c}{\omega} - \frac{\omega}{\omega_c} \right) \ln \frac{1 + \omega/\omega_c}{1 - \omega/\omega_c} \right].$$

For $\omega = 0$, the latter expression approaches the limit $\ln A(0) = 4\omega_c\tau/\pi$, so that

$$\ln A(\omega)/A(0) = -\frac{2\omega_c\tau}{\pi} \left[1 + \frac{1}{2} \left(\frac{\omega_c}{\omega} - \frac{\omega}{\omega_c} \right) \ln \frac{1 + \omega/\omega_c}{1 - \omega/\omega_c} \right]. \quad (1.08)$$

which is the attenuation characteristic shown in Fig. 3.

Other attenuation characteristics with a linear phase characteristic between $\omega/\omega_c = 0$ and 1 are possible with other types of variations in the attenuation or phase characteristic for $\omega/\omega_c > 1$ than assumed above. For example, the attenuation characteristics may be assumed

constant for $\omega/\omega_c > 1$, in which case the attenuation characteristic will be somewhat different for $\omega/\omega_c < 1$ and the phase characteristic different for $\omega/\omega_c > 1$, as illustrated in Fig. 3. (The solution for the latter case is given in Reference 2.) It will be noticed that there is a comparatively minor difference between the attenuation characteristics for $\omega/\omega_c < 1$ in the above cases, so that the attenuation characteristic for $\omega/\omega_c > 1$ has a relatively minor effect, provided there is no discontinuity near $\omega/\omega_c = 1$. The transmission loss characteristics shown in Fig. 3 represent a close approximation to the type of characteristic employed in pulse transmission systems, as will be shown later.

In the above examples low-pass characteristics were assumed. For high-pass characteristics the algebraic sign of the phase is reversed with respect to the amplitude characteristic as indicated in Fig. 4, which also illustrates relationships for band-pass characteristics. The band-pass characteristics are obtained by connecting low-pass and high-pass networks in tandem. The resultant attenuation and phase characteristics are obtained by adding the low and high-pass attenuation and phase characteristics, as illustrated in the figure. In the second case shown in the figure, the band-pass characteristic is assumed to have a linear phase characteristic in the transmission band, in which case the attenuation characteristic will not be symmetrical about the midband frequency, unless the latter is high in relation to the bandwidth. The third case illustrates the type of band-pass characteristic encountered in wire systems with a low-frequency cutoff. There will then be phase distortion at the low end of the band, since it is not feasible with a fairly sharp low-frequency cutoff to obtain a linear phase characteristic in the transmission band.

If the amplitude or attenuation characteristic of a transmission system is modified, it will be accompanied by a modification in the phase characteristic. Of basic importance are cosine modifications in the attenuation and amplitude characteristics. Let the modified amplitude characteristic be of the form

$$A(\omega) = A_0(\omega) e^{a \cos \omega \tau}, \quad (1.09)$$

where $A_0(\omega)$ is the original amplitude characteristic. The modified attenuation characteristic is then

$$\ell n A(\omega) = \ell n A_0(\omega) + a \cos \omega \tau. \quad (1.10)$$

In accordance with (1.05) the modified phase characteristic becomes,

$$\begin{aligned} \psi^0(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\ell n A_0(u)}{\omega - u} du + \frac{a}{\pi} \int_{-\infty}^{\infty} \frac{\cos u \tau}{\omega - u} du, \\ &= \psi_0(\omega) + a \sin \omega \tau, \end{aligned} \quad (1.11)$$

where $\psi_0(\omega)$ is the phase characteristic of the original amplitude characteristic $A_0(\omega)$.

Thus, for any cosine modification in the attenuation characteristic there is a corresponding sine modification in the phase characteristic, and for any sine modification in the phase characteristic a corresponding cosine modification in the attenuation characteristic. In general any modification in the attenuation characteristic may be represented by a Fourier cosine series, in which case the modification in the phase characteristic will be the corresponding Fourier sine series.

With a cosine modification in the amplitude rather than in the at-

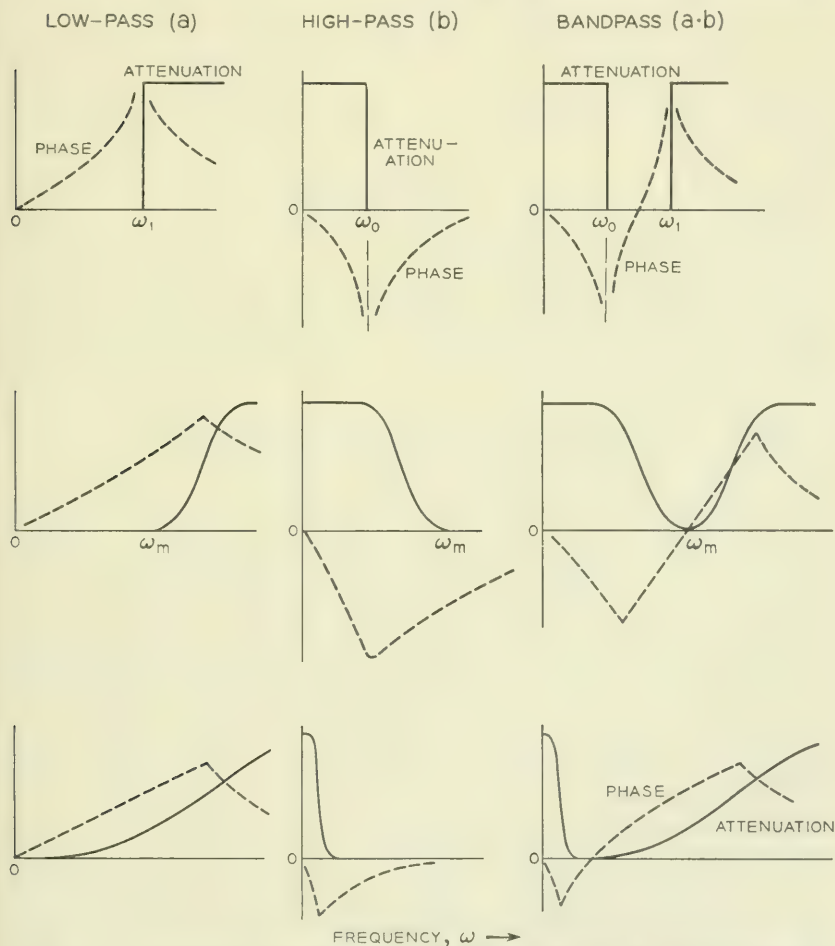


Fig. 4 — Attenuation and phase shift for various types of transmission frequency characteristics.

tenuation characteristic

$$A(\omega) = A_0(\omega) [1 + a \cos \omega \tau], \quad (1.12)$$

and the corresponding phase characteristic becomes

$$\begin{aligned} \psi^0(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\ell n A_0(u) [1 + a \cos u \tau]}{\omega - u} du, \\ &= \psi_0(\omega) + 2 \tan^{-1} \frac{r \sin \omega \tau}{1 + r \cos \omega \tau}, \\ &= \psi_0(\omega) + 2[r \sin \omega \tau + \frac{r^2}{2} \sin 2 \omega \tau \\ &\quad + \frac{r^3}{3} \sin 3 \omega \tau + \dots] \end{aligned} \quad (1.13)$$

where

$$r = \frac{1}{a} [1 \mp \sqrt{1 - a^2}], \quad (1.14)$$

and the minus sign is to be used.

Thus, a cosine modification in the amplitude characteristic is accompanied by an infinite series of sine deviations in the phase characteristic. For sufficiently small values of a , $r \cong a/2$ and (1.13) reduces to (1.11).

2. FREQUENCY AND IMPULSE TRANSMISSION CHARACTERISTICS

In dealing with pulse transmission, it is customary to consider three basic types of time variations of currents and electromotive forces, a cisoidal variation, a unit impulse and a unit step. The cisoidal variation, $e^{i\omega t}$, is basic in the solution of network and transmission problems in terms of complex impedances and admittances. The unit impulse is a current or electromotive force of very high intensity and short duration, such that the area under the impulse is unity. The unit step is a current or electromotive force which is zero for $t < 0$ and unity thereafter.

The time responses of networks or transmission systems to these three basic time functions are interrelated so that each may be obtained when one of the others is known. Furthermore, the time responses for electromotive forces or currents of arbitrary wave shape may be obtained from the response characteristic for any one of these basic time functions.

The pulses applied in pulse systems can usually be approximated by impulses. Furthermore, with impulses certain simple relationships can be established which are either obscured or more complicated when a

unit step is assumed. For these reasons, only the transmission characteristic for impulses will be considered here, or for pulses of sufficiently short duration to be regarded as impulses.

Corresponding to any transmission-frequency characteristic is an impulse transmission characteristic, $P(t)$, which designates the received pulse as a function of time for a transmitted unit impulse. The impulse and transmission frequency characteristics are interrelated by the following Fourier integral relations

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T(i\omega) e^{i\omega t} d\omega, \quad (2.01)$$

$$T(i\omega) = \int_{-\infty}^{\infty} P(t) e^{-i\omega t} dt. \quad (2.02)$$

The transmission characteristic for an applied pulse or signal of arbitrary shape $G(t)$ is given by

$$H(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T(i\omega) S(i\omega) e^{i\omega t} d\omega, \quad (2.03)$$

where $S(i\omega)$ is the frequency spectrum of the applied pulse and is given by

$$S(i\omega) = \int_{-\infty}^{\infty} G(t) e^{-i\omega t} dt. \quad (2.04)$$

In the case of a symmetrical pulse $S(i\omega)$ is a real function.

In view of (1.01), expression (2.03) may also be written

$$H(t) = \frac{1}{\pi} \int_0^{\infty} A(\omega) S(\omega) \cos [\omega t - \psi(\omega)] d\omega, \quad (2.05)$$

where the relations $A(-\omega) = A(\omega)$, $S(-\omega) = S(\omega)$, $\psi(-\omega) = -\psi(\omega)$ have been used, and it is assumed that $S(i\omega) = S(\omega)$ is a real function, as for a symmetrical pulse.

In most pulse transmission systems, the applied pulses can be approximated by short rectangular pulses. Rectangular pulses of unit amplitude and duration δ have a frequency spectrum

$$S(\omega) = \delta \frac{\sin \omega\delta/2}{\omega\delta/2}. \quad (2.06)$$

The same pulse transmission characteristic as when an impulse is applied is obtained with a rectangular pulse if $A(\omega)$ is modified by the factor $(\omega\delta/2)/\sin(\omega\delta/2)$. In the following it will be assumed that the applied pulses are of sufficiently short duration to be regarded as im-

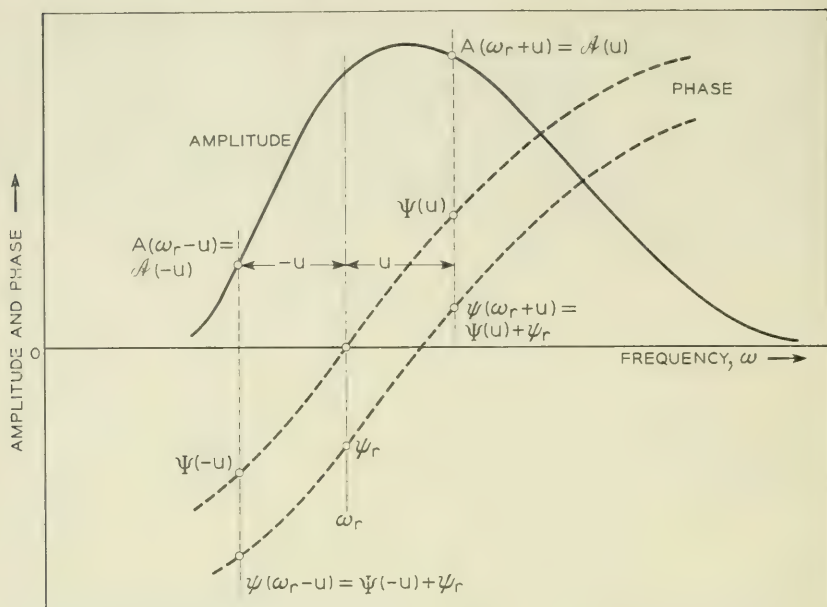


Fig. 5 — Transfer of reference frequency from $\omega = 0$ to $\omega = \omega_r$.

pulses or that otherwise the above modification is applied, in which case

$$P(t) = \frac{\delta}{\pi} \int_0^{\infty} A(\omega) \cos [\omega t - \psi(\omega)] d\omega. \quad (2.07)$$

In the latter equation $A(\omega)$ can also be regarded as the frequency spectrum of a pulse applied to a transmission system having a constant amplitude characteristic and a phase characteristic $\psi(\omega)$ over the band of the pulse spectrum.

Equation (2.07) applies to any type of transmission-frequency characteristic and is convenient in this form for low-pass characteristics. For band-pass characteristics as shown in Fig. 5 however, it is convenient from the standpoint of general analysis as well as for numerical evaluation to use a reference frequency ω_r within the transmission band, that is, to employ the transformation $\omega = \omega_r + u$ $d\omega = du$.

With the notation

$$\begin{aligned} \alpha(u) &= A(\omega) = A(u + \omega_r), \\ \Psi(u) &= \psi(\omega) - \psi(\omega_r) = \psi(\omega) - \psi_r, \end{aligned} \quad (2.08)$$

equation (2.07) can be written:

$$P(t) = \cos(\omega_r t - \psi_r)[R_-(t) + R_+(t)] \\ + \sin(\omega_r t - \psi_r)[Q_-(t) - Q_+(t)]. \quad (2.09)$$

$$R_- = \frac{\delta}{\pi} \int_0^{\omega_r} \alpha(-u) \cos[ut + \Psi(-u)] du, \quad (2.10)$$

$$R_+ = \frac{\delta}{\pi} \int_0^{\infty} \alpha(u) \cos[ut - \Psi(u)] du,$$

$$Q_- = \frac{\delta}{\pi} \int_0^{\omega_r} \alpha(-u) \sin[ut + \Psi(-u)] du, \quad \text{and} \quad (2.11)$$

$$Q_+ = \frac{\delta}{\pi} \int_0^{\infty} \alpha(u) \sin[ut - \Psi(u)] du.$$

The envelope $\bar{P}(t)$ of the impulse transmission characteristic is given by

$$\bar{P}(t) = [(R_- + R_+)^2 + (Q_- - Q_+)^2]^{1/2}. \quad (2.12)$$

Comparison of (2.09) with (2.07) shows that R_- and R_+ can be identified with the impulse characteristics of low-pass systems having the same frequency characteristics as the bandpass system below and above ω_r . The impulse characteristics Q_- and Q_+ which arise from asymmetry in the transmission characteristic with respect to ω_r are not present in low-pass systems, since by definition the amplitude characteristic has even symmetry and the phase characteristic odd symmetry with respect to zero frequency.

The first and second components of (2.09) are referred to as the in-phase and quadrature components of the impulse characteristic of band-pass systems.³ The transmission-frequency characteristic may correspondingly be regarded as made up of a component with even symmetry and another component with odd symmetry about ω_r , as indicated in Fig. 6. These two components, together with the in-phase and quadrature components, will depend on the choice of ω_r . However, $P(t)$ as given by (2.09) and the envelope as given by (2.12), will remain the same, since a single impulse characteristic is associated with a given transmission-frequency characteristic.

With the customary pulse transmission methods, the reference frequency ω_r may be identified with a modulating or carrier frequency, which has a special significance when the envelope of a sequence of received pulses is considered. Although for a single pulse the envelope

is always the same, for a sequence of pulses the resultant envelope of the received pulse train will depend on the in-phase and quadrature components.⁴ The reason for this is that one has even and the other odd symmetry about the peak amplitude of the envelope for a single pulse, when the phase characteristic is linear.

In order to compare the transmission performance as the reference or carrier frequency is changed, it is necessary to determine the in-phase and quadrature components for each carrier frequency under considera-

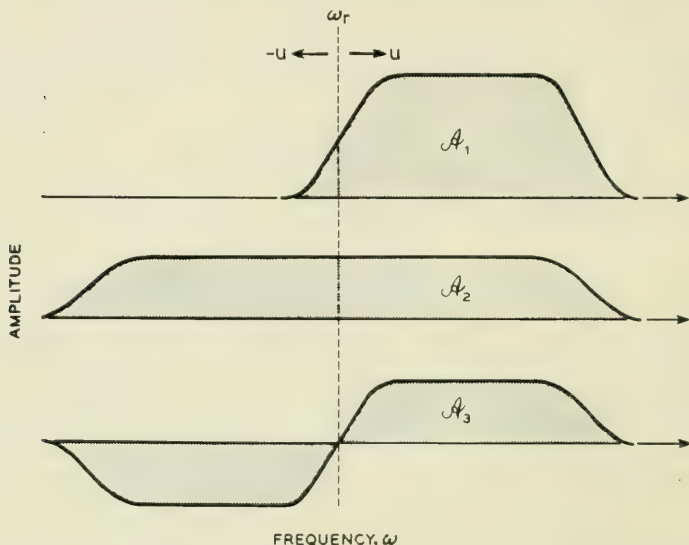


Fig. 6 — Decomposition of amplitude characteristic A_1 asymmetrical with respect to ω_r into a component A_2 of even symmetry and a component A_3 of odd symmetry about ω_r . When the phase shift is linear, $A_1 = A_2 + A_3$.

tion. One method is to evaluate integrals (2.10) and (2.11) for each carrier frequency, which may be facilitated by resolving the transmission-frequency characteristic into symmetrical and anti-symmetrical components as indicated in Fig. 6. This, however, is a rather elaborate procedure which can be avoided with the aid of a simple translation from one reference or carrier frequency to another, as shown below, provided the in-phase and quadrature components or the envelope has been determined for one reference frequency.

Equation (2.09) may also be written, with $\varphi = \varphi(t)$:

$$\begin{aligned} P(t) &= \cos(\omega_r t - \psi_r - \varphi) \bar{P}(t), \\ &= \cos(\omega_r t - \psi_r) \cos \varphi \bar{P}(t) + \sin(\omega_r t - \psi_r) \sin \varphi \bar{P}(t). \end{aligned} \quad (2.13)$$

Comparison of (2.13) with (2.09) shows that:

$$\begin{aligned} R_- + R_+ &= \cos \varphi \bar{P}(t), \\ Q_- - Q_+ &= \sin \varphi \bar{P}(t), \end{aligned} \quad (2.14)$$

$$\tan \varphi = (Q_- - Q_+) / (R_- + R_+). \quad (2.15)$$

To find the corresponding components when ω_r is changed to ω_r' , equation (2.13) may be written

$$\begin{aligned} P(t) &= \cos[\omega_r' t - \psi_r' - (\omega_r' - \omega_r)t + (\psi_r' - \psi_r) - \varphi] \bar{P}(t) \\ &= \cos(\omega_r' t - \psi_r' - \varphi') \bar{P}(t), \end{aligned} \quad (2.16)$$

where $\varphi' = \varphi'(t)$ is given by:

$$\begin{aligned} \varphi' &= \varphi + (\omega_r' - \omega_r)t - (\psi_r' - \psi_r), \\ &= \varphi + \omega_y t - \psi_y. \end{aligned} \quad (2.17)$$

Thus, when the reference frequency is changed by ω_y and its phase by ψ_y , the corresponding in-phase and quadrature components become:

$$\begin{aligned} R_-' + R_+' &= \cos(\varphi + \omega_y t - \psi_y) \bar{P}(t), \quad \text{and} \\ Q_-' - Q_+' &= \sin(\varphi + \omega_y t - \psi_y) \bar{P}(t) \end{aligned} \quad (2.18)$$

To summarize, when the in-phase and quadrature components have been determined for any reference frequency ω_r from (2.10) and (2.11), and the envelope $\bar{P}$ together with the function φ from (2.12) and (2.14), the in-phase and quadrature components for another reference frequency ω_r' can readily be determined with the aid of (2.18). In the particular case where the amplitude characteristic has even and the phase characteristic odd symmetry with respect to the midband frequency, the quadrature component disappears with respect to the midband frequency, so that $\varphi = 0$ and (2.18) simplifies to

$$\begin{aligned} R_-' + R_+' &= \cos(\omega_y t - \psi_y) \bar{P}(t), \quad \text{and} \\ Q_-' - Q_+' &= \sin(\omega_y t - \psi_y) \bar{P}(t). \end{aligned} \quad (2.19)$$

The above relations (2.18) and (2.19) facilitate comparison of transmission performance as the reference or carrier frequency is changed, for example the comparison of double with vestigial sideband transmission, as illustrated in section 14.

3. IDEALIZED CHARACTERISTICS WITH SHARP CUTOFF

In pulse transmission theory, particularly in dealing with transmission capacity of idealized transmission systems, an ideal low-pass transmission frequency characteristic is ordinarily assumed, with constant amplitude and delay in the transmission band together with an abrupt cutoff at the top frequency and zero amplitude beyond, as shown in Fig. 7. As is evident from Fig. 1, this type of characteristic is an abstraction which cannot be physically realized since it will have phase distortion and infinite transmission delay. It can, however, be approached with sufficiently elaborate phase equalization.

For the above type of characteristic, $A(\omega) = 1$ between $\omega = 0$ and ω_1 , while $\psi(\omega) = \omega\tau_d$, where τ_d is the transmission delay. With these values in (2.07):

$$P(t) = \frac{\delta\omega_1}{\pi} \frac{\sin \omega_1 t_0}{\omega_1 t_0}, \quad (3.01)$$

where $t_0 = t - \tau_d$ is the time referred to the peak amplitude of the received pulse.

The resultant pulse transmission characteristic is shown in Fig. 7, with the factor $\delta\omega_1/\pi$ omitted. The peak amplitude is attained after an infinite time, since the above type of characteristic can be realized only with $\tau_d \rightarrow \infty$. The impulse characteristic is zero when $\omega_1 t_0 = \pm n\pi$, or $t_0 = \pm\tau_1, \pm 2\tau_1, \dots \pm n\tau_1$ where

$$\tau_1 = \frac{1}{2f_1}. \quad (3.02)$$

Impulses can thus be transmitted at the latter intervals without

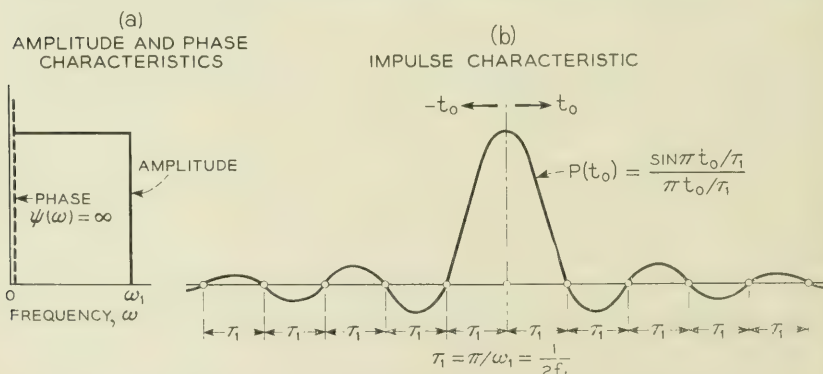


Fig. 7 — Idealized low-pass characteristic with sharp complete cut-off.

mutual interference between the peaks of the received pulses. This is a basic theorem underlying the determination of the transmission capacity of idealized systems.³

For an idealized bandpass characteristic between ω_0 and ω_1 , it follows from (2.09) with $\Psi(u) = u\tau_d$ and $\Psi(-u) = -u\tau_d$ that the impulse characteristic with respect to the midband frequency $\omega_r = \omega_m$ is

$$P(t) = 2 \cos[\omega_m t_0 - \psi_0] \bar{P}(t), \quad (3.03)$$

where $\bar{P}(t)$ is given by (3.01) and $\psi_0 = \Psi_m - \omega_m \tau_d$ is the phase intercept at zero frequency. For the transmission characteristic to be ideal in the sense that the peak pulse amplitude occurs when $t_0 = t - \tau_d = 0$, it is necessary that $\psi_0 = \pm n\pi$, where n is an integer. This is not necessary if the bandwidth is small in relation to the midband frequency. There will then be a large number of cycles of the modulating frequency ω_m within the envelope $\bar{P}(t)$, and the latter can be recovered by envelope detection regardless of the phase of the modulating frequency.

With $\psi_0 = \pm n\pi$,

$$P(t) = \frac{2\omega_s \delta}{\pi} \cos \omega_m t_0 \frac{\sin \omega_s t_0}{\omega t_0}, \quad (3.04)$$

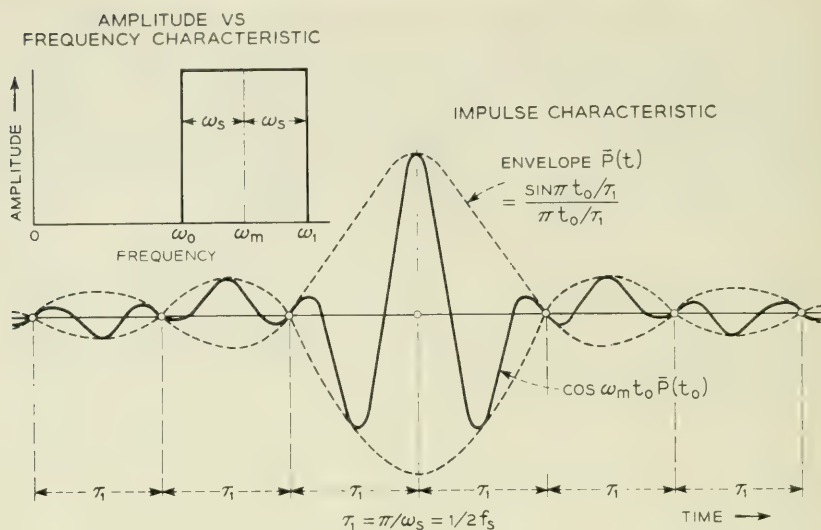
$$= \frac{\omega_1 \delta}{\pi} \frac{\sin \omega_1 t_0}{\omega_1 t_0} - \frac{\omega_0 \delta}{\pi} \frac{\sin \omega_0 t_0}{\omega_0 t_0}, \quad (3.05)$$

where $\omega_m = (\omega_0 + \omega_1)/2$ and $\omega_s = (\omega_1 - \omega_0)/2$.

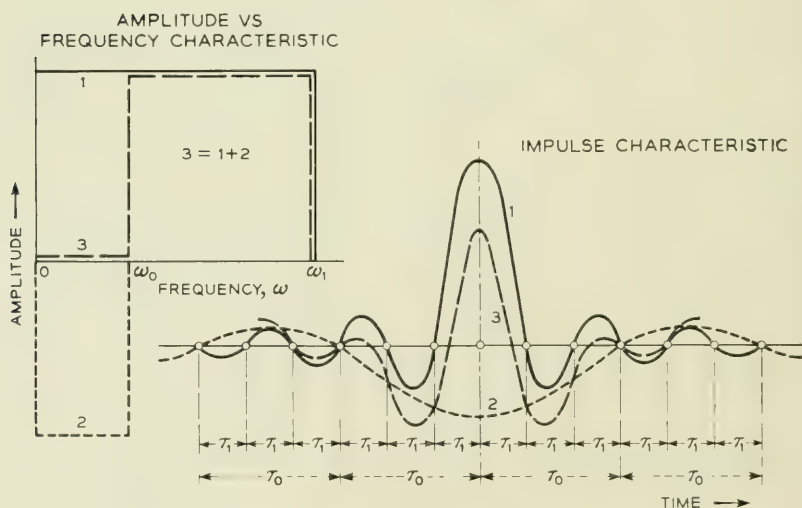
The shape of the impulse characteristic as given by (3.04) is illustrated in the upper half of Fig. 8. Alternately the impulse characteristic may be regarded as made up of two components in accordance with (3.05). The first component corresponds to a low-pass characteristic of bandwidth ω_1 , the second component to a negative low-pass characteristic of bandwidth ω_0 , as indicated in the lower part of the Fig. 8.

The factor $\sin \omega_s t_0 / \omega_s t_0$ in (3.04) is zero at the same intervals as for a low-pass characteristic of bandwidth ω_s , as shown in Fig. 8, so that pulses may be transmitted at the same rate without mutual interference between pulse peaks. The bandwidth in the present case, however, is $2\omega_s = \omega_1 - \omega_0$, so that for the same bandwidth the pulse transmission rate is half as great as for a low-pass characteristic.

An exception to this is the particular case when $\omega_1 = 2\omega_0$, so that the total bandwidth is ω_0 . The factor $\sin \omega_0 t_0 / \omega_0 t_0$ in (3.05) is then zero at intervals $\tau_0 = 1/2f_0$, while the factor $\sin \omega_1 t_0 / \omega_1 t_0$ is zero at intervals $1/2f_1 = 1/4f_0$, as shown in Fig. 9. Pulses may accordingly in principle be



(a) REPRESENTATION OF IMPULSE CHARACTERISTIC AS ENVELOPE MODULATED BY MIDBAND FREQUENCY



(b) REPRESENTATION OF AMPLITUDE VS FREQUENCY AND IMPULSE CHARACTERISTICS, 3, AS THE SUM OF A POSITIVE LOW-PASS CHARACTERISTIC, 1, AND A NEGATIVE LOW-PASS CHARACTERISTIC, 2

Fig. 8 — Idealized band-pass characteristics and corresponding impulse transmission characteristics.

transmitted without mutual interference at the same rate as for a low-pass characteristic of bandwidth ω_0 , or at the same rate as with single sideband transmission over a band-pass system of bandwidth ω_0 . More generally, pulses can in principle be transmitted without mutual interference between pulse peaks at the same rate as for a low-pass characteristic of bandwidth $\omega_1 - \omega_0 = 2\omega_s$ if ω_0 is a multiple of $\omega_1 - \omega_0$. It should be noted however, that this pulse transmission rate cannot actually be realized since the phase characteristic will have infinite slope, so that the transmission delay will be infinite. In addition, the zero frequency phase intercept ψ_0 must be $\pm n\pi$, a condition which cannot be attained or remain stable in view of the infinite slope of the phase characteristic.

With the envelope given by the factor $\sin \omega_s t_0 / \omega_s t_0$ in (3.04), the in-phase and quadrature components for any reference frequency can be determined with the aid of (2.19). If the lower band-edge is selected, i.e. $\omega_r = \omega_0$, then $\omega_y = \omega_s$. With a linear phase characteristic $\psi_y = \omega \tau_d$, so that in (2.19) $\omega_y t - \psi_y = \omega_s t_0$. The in-phase and quadrature components are accordingly obtained by multiplying the envelope by $\cos \omega_s t_0$ and $\sin \omega_s t_0$, respectively.

As an alternate method, the two components can be obtained from

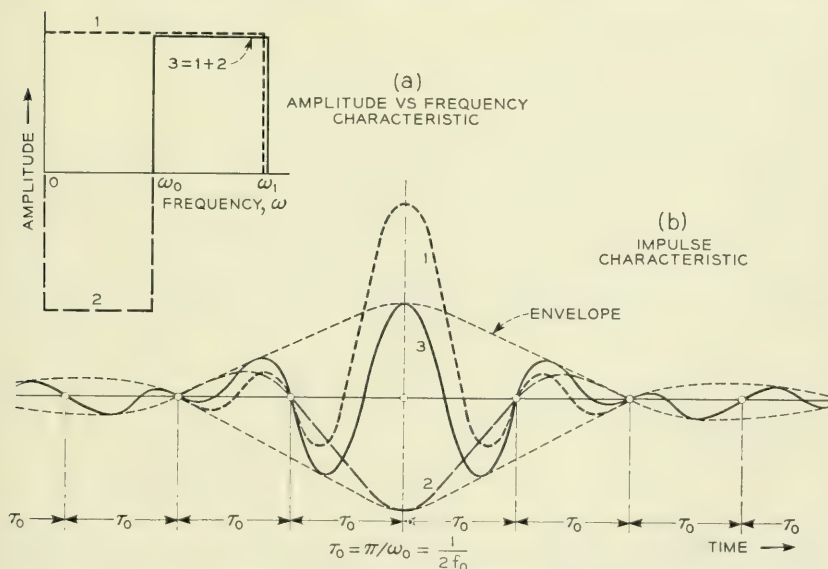


Fig. 9 — Special case of idealized band-pass characteristic in which $\omega_1 = 2\omega_0$ and resultant impulse characteristic is zero at intervals $\tau_0 = \frac{1}{2f_0}$.

(2.09), which with $R_- = 0$, $Q_- = 0$ becomes:

$$P(t) = \cos \omega_0 t_0 R_+(t) + \sin \omega_0 t_0 Q_+(t), \quad (3.06)$$

with

$$R_+ = \frac{\delta}{\pi} \int_0^{\omega_b} \cos ut_0 du, \quad (3.07)$$

$$= \frac{\hat{\epsilon} \omega_b}{\pi} \frac{\sin \omega_b t_0}{\omega_b t_0} = \frac{2\hat{\epsilon} \omega_s}{\pi} \cos \omega_s t_0 \frac{\sin \omega_s t_0}{\omega_s t_0}, \quad \text{and}$$

$$Q_+ = \frac{\delta}{\pi} \int_0^{\omega_b} \sin ut_0 du, \quad (3.08)$$

$$= \frac{\hat{\epsilon} \omega_b}{\pi} \frac{1 - \cos \omega_b t_0}{\omega_b t_0} = \frac{2\hat{\epsilon} \omega_s}{\pi} \sin \omega_s t_0 \frac{\sin \omega_s t_0}{\omega_s t_0},$$

where $\omega_b = 2\omega_s$ is the bandwidth. It will be noticed that R_+ and Q_+ are obtained by multiplying the envelope by $\cos \omega_s t_0$ and $\sin \omega_s t_0$ in accordance with (2.19).

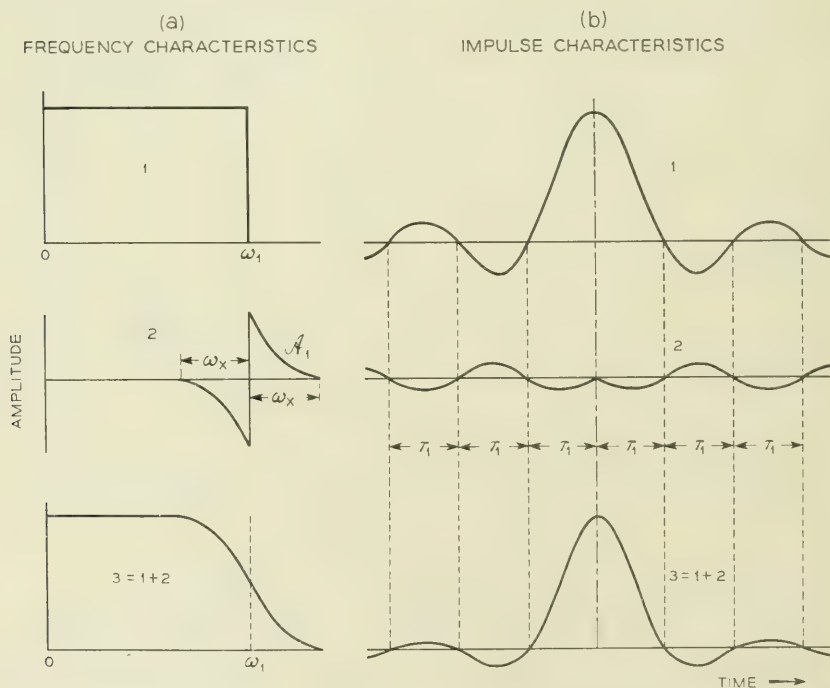


Fig. 10 — Idealized transmission characteristic with gradual cut-off, 3, obtained by superposition of characteristic with sharp cut-off, 1, and characteristic, 2, with odd symmetry about ω_1 . Linear phase shift assumed.

4. IDEALIZED CHARACTERISTICS WITH GRADUAL CUTOFF

The idealized transmission characteristics discussed above are of principal interest in that they indicate the physical limitations on pulse transmission rates for a given bandwidth. Even if these impulse characteristics could be realized without undue difficulties from the standpoint of phase equalization, they would be impracticable in most applications. Their oscillatory nature would entail the use of discrete pulse positions and precise synchronized sampling at fixed intervals, and would preclude certain methods of pulse modulation and detection.

The non-linearity in the phase characteristic as well as the oscillations in the impulse characteristic can be reduced with a gradual rather than a sharp cut-off, as illustrated in Fig. 10. It is assumed that an ideal characteristic with a sharp cutoff is supplemented by an amplitude characteristic α_1 which has odd symmetry about the cutoff frequency ω_1 , i.e., $\alpha_1(-u) = -\alpha_1(u)$.

If the latter component alone is considered, and a linear phase characteristic assumed, it follows from (2.09) with $\omega_1 = \omega_r$ that the effect of this component on the pulse transmission characteristic is given by

$$P_1(t) = -Q_1 \sin \omega_1 t_0, \quad (4.01)$$

where $t_0 = t - \tau_d$ and

$$Q_1 = \frac{2\delta}{\pi} \int_0^{\omega_x} \alpha_1(u) \sin ut_0 du. \quad (4.02)$$

The function $P_1(t)$ will be zero at the same points as the original pulse transmission characteristic with a sharp cut-off at ω_1 and under certain conditions also at other points. It will modify the original impulse characteristic by reducing the oscillatory tail, as illustrated in Fig. 10, but the zero points remain unchanged.³

With the above modification, the resultant impulse characteristic obtained by superposition of (3.01) and (4.02) becomes

$$\begin{aligned} P(t) &= \frac{\delta}{\pi} \sin \omega_1 t_0 \left(\frac{1}{t_0} - 2 \int_0^{\omega_x} \alpha_1(u) \sin ut_0 du \right), \\ &= \frac{\delta}{\pi} \sin \omega_1 t_0 F(t), \end{aligned} \quad (4.03)$$

where

$$F(t) = \left[\frac{1}{t_0} - 2 \int_0^{\omega_x} \alpha_1(u) \sin ut_0 du \right]. \quad (4.04)$$

In the following the expression for $F(t)$ is given for the case when the

band-edge is modified by a supplementary characteristic of the form

$$\begin{aligned}\alpha_1(u) &= \frac{1}{2}(1 - \sin \pi u/2\omega_x) & u < \omega_x, \\ &= 0 & u > \omega_x.\end{aligned}\quad (4.05)$$

This form of α_1 represents a close approximation to actual modifications of band-edges by a gradual cutoff and also results in rather simple expressions for the modified impulse characteristic

With (4.05) in (4.04),

$$\begin{aligned}F(t) &= \frac{1}{t_0} - \int_0^{\omega_x} (1 - \sin \pi u/2\omega_x) \sin ut_0 \, du, \\ &= \frac{1}{t_0} - \omega_x \left[\frac{1 - \cos \omega_x t_0}{\omega_x t_0} + \frac{\cos \omega_x t_0}{\pi + 2\omega_x t_0} - \frac{\cos \omega_x t_0}{\pi - 2\omega_x t_0} \right], \\ &= \omega_x \cos \omega_x t_0 \left[\frac{1}{\omega_x t_0} + \frac{1}{\pi - 2\omega_x t_0} - \frac{1}{\pi + 2\omega_x t_0} \right], \\ &= \frac{1}{t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2}.\end{aligned}\quad (4.06)$$

The impulse characteristic obtained from (4.03) is

$$P(t) = \frac{\delta\omega_1}{\pi} \frac{\sin \omega_1 t_0}{\omega_1 t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2} \quad (4.07)$$

For the particular case shown in Fig. 11 the value of ω_x is taken to be $\omega_1/2$.

For a symmetrical bandpass characteristic, as shown in Fig. 12,

$$P(t) = 2 \cos (\omega_m t_0 - \psi_0) \bar{P}(t). \quad (4.08)$$

$\bar{P}(t)$ is obtained by replacing ω_1 by ω_s in (4.07), and ψ_0 is the phase intercept at zero frequency as in connection with (3.03). This gives

$$\bar{P}(t) = \frac{\delta\omega_s}{\pi} \frac{\sin \omega_s t_0}{\omega_s t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2}. \quad (4.09)$$

For the particular case shown in Fig. 12, the value of ω_x is taken to be $\omega_s/2$.

The in-phase and quadrature components with respect to any frequency are obtained from (2.19) with $\psi_y = \omega\tau_d$ and are shown in Fig. 12 for the particular case in which the reference frequency is displaced from the midband frequency by $\omega_y = \omega_s$.

5. IDEALIZED CHARACTERISTICS WITH NATURAL LINEAR PHASE SHIFT

With the type of amplitude characteristics discussed above it is necessary to employ phase equalization to obtain a linear phase characteristic. Furthermore, oscillations of appreciable amplitude remain in the impulse characteristic. A virtually linear phase characteristic together with a reduction of these oscillations can be attained by a further extension of the gradual cut-off in Fig. 10, such that $\omega_x = \omega_1$. An amplitude characteristic of this type, together with the corresponding impulse

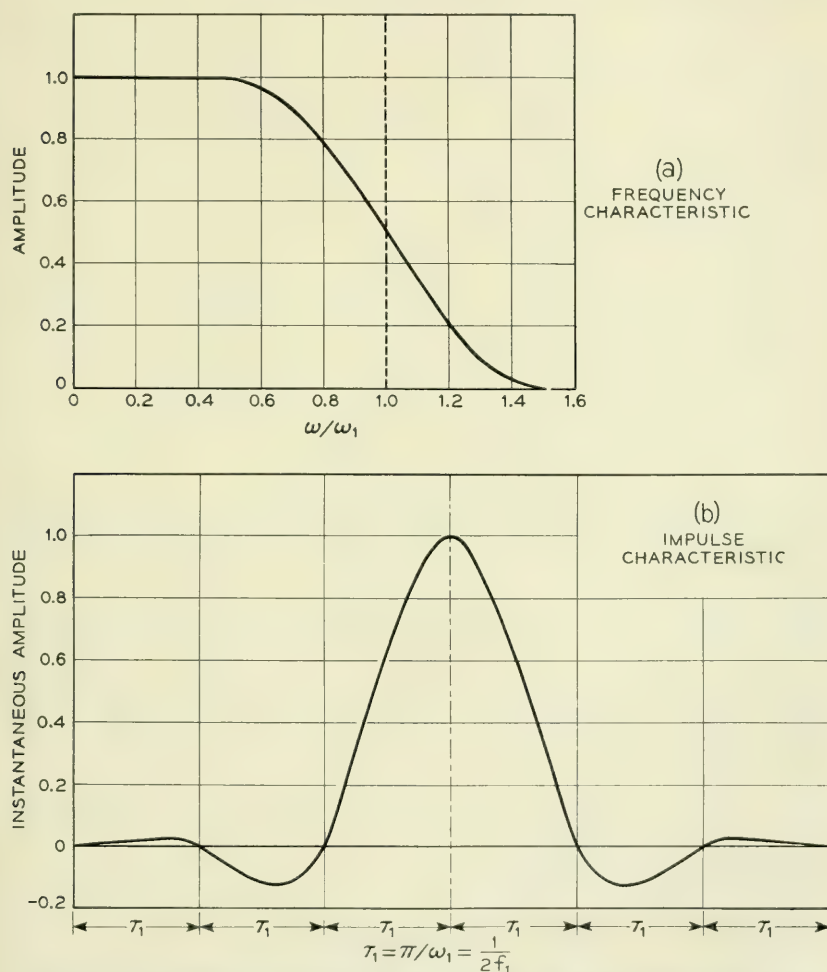
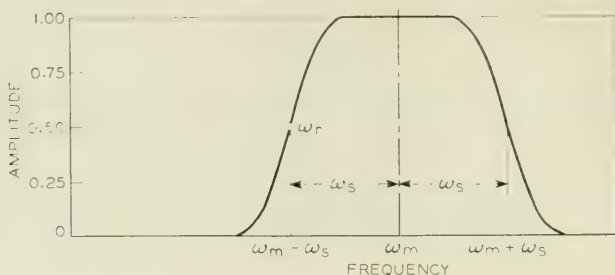
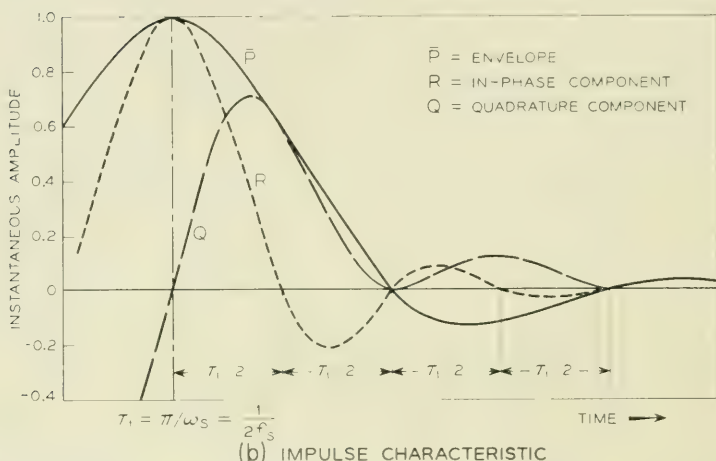


Fig. 11 — Low-pass characteristic with gradual cut-off and associated impulse characteristic. Linear phase characteristic assumed.



(a) FREQUENCY CHARACTERISTIC



(b) IMPULSE CHARACTERISTIC

Fig. 12—Symmetrical band-pass characteristic with gradual cut-off and associated impulse characteristic. In-phase and quadrature components shown with respect to $\omega_r = \omega_m - \omega_s$.

characteristic is shown in Fig. 13. The supplementary amplitude characteristic and the impulse characteristic are obtained by making $\omega_x = \omega_1$ in (4.05) and 4.07).

The resultant amplitude characteristic between $\omega = 0$ and $\omega = 2\omega_1$ in this case becomes

$$A(\omega) = \frac{1}{2} \left[1 + \cos \frac{\pi \omega}{2\omega_1} \right] = \cos^2 \frac{\pi \omega}{4\omega_1}, \quad (5.01)$$

and the impulse characteristic:

$$P(t) = \frac{\epsilon \omega_1}{\pi} \frac{\sin 2\omega_1 t_0}{2\omega_1 t_0 [1 - (2\omega_1 t_0 / \pi)^2]}, \quad (5.02)$$

where ω_1 is the bandwidth to the half-amplitude point on the trans-

mission frequency characteristic and $2\omega_1$ the bandwidth to the point of zero amplitude.

In Fig. 13 is also shown the amplitude characteristic given by (1.08), which will have a linear phase characteristic in the transmission band, i.e. from $\omega = 0$ to $2\omega_1$. Because of the close approximation of (5.01) to the proper type of amplitude characteristic as regards phase linearity, the phase characteristic associated with (5.01) may for practical purposes be regarded as linear.

For a symmetrical band-pass characteristic as shown in Fig. 14, the impulse characteristic is given by (4.08) and the envelope by (4.09) with $\omega_x = \omega_s$, or

$$\bar{P}(t) = \frac{\epsilon\omega_s}{\pi} \frac{\sin 2\omega_s t_0}{2\omega_s t_0 [1 - (\omega_s t_0/\pi)^2]}. \quad (5.03)$$

The in-phase and quadrature components shown in Fig. 14 with

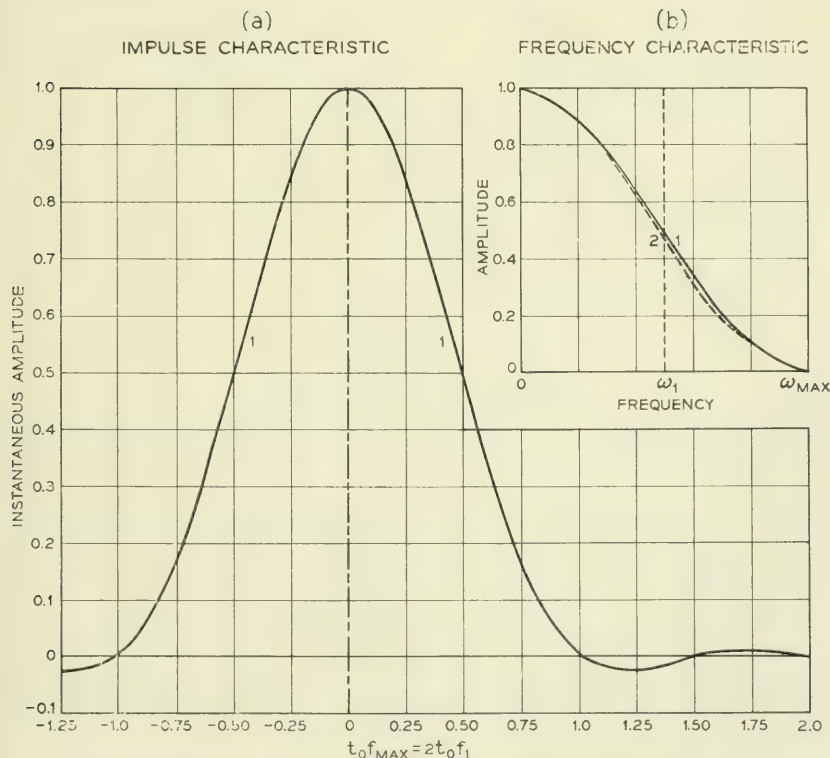


Fig. 13 — Low-pass transmission frequency characteristic, 1, and associated impulse characteristic. Frequency characteristic, 2, is same as shown by solid lines in Fig. 3 and has a linear phase characteristic between $\omega = 0$ and ω_{max} .

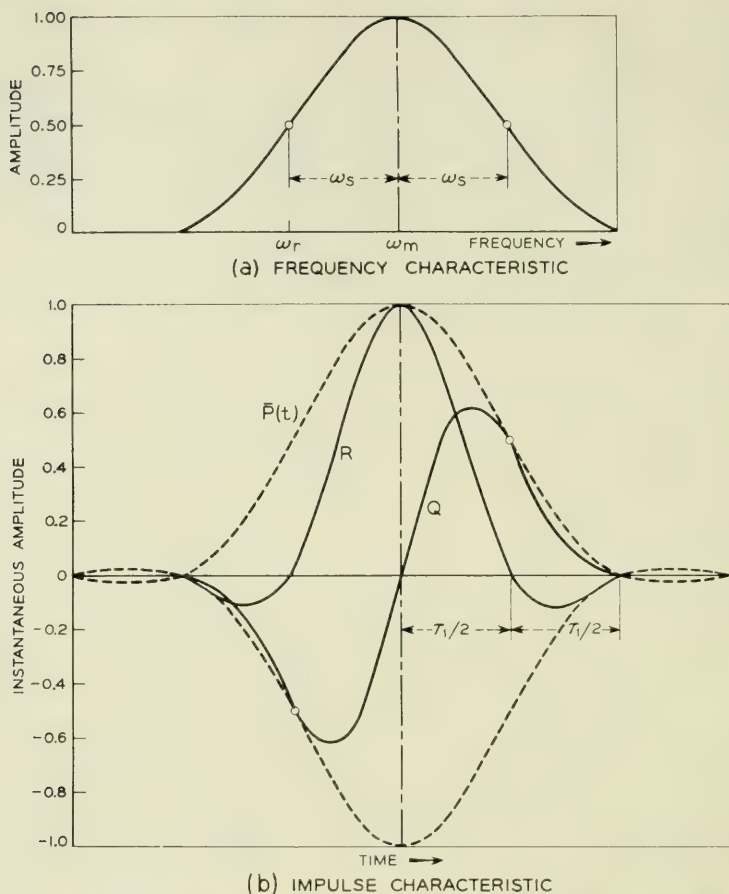


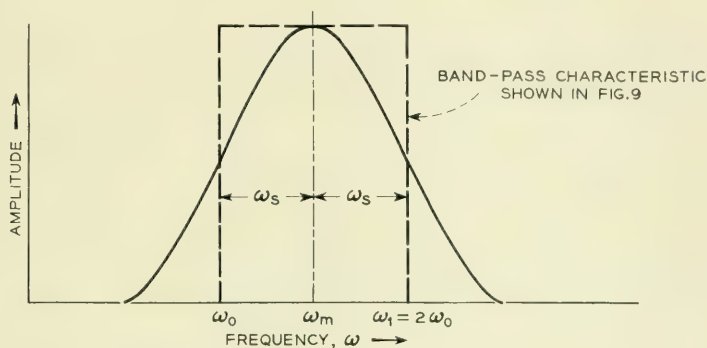
Fig. 14 — Symmetrical band-pass characteristic with linear phase shift and corresponding impulse characteristic. In-phase and quadrature components shown with respect to $\omega_r = \omega_m - \omega_s$.

respect to a reference frequency at the midpoint of the band-edge are obtained from (2.19) with $\omega_y = \omega_s$. This gives $\bar{P}(t) \cos \omega_s t_0$ for the in-phase and $\bar{P} \sin \omega_s t_0$ for the quadrature component.

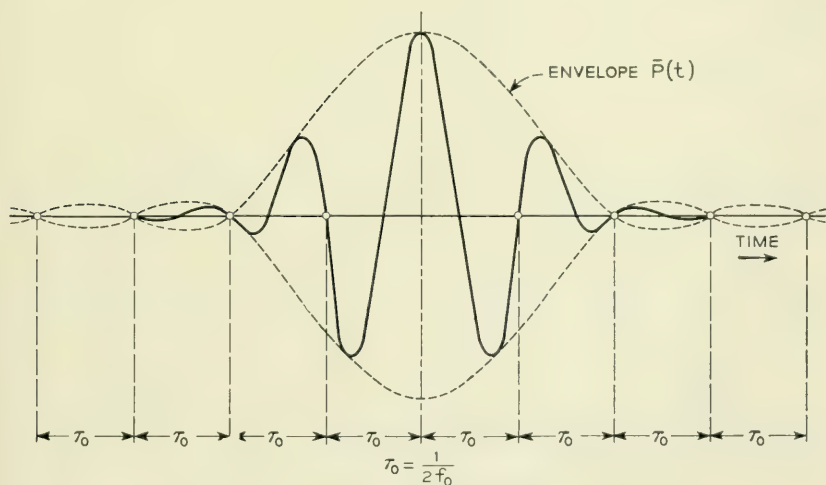
In Fig. 15 is shown a special case of a band-pass characteristic, which corresponds to that illustrated in Fig. 9 with $\omega_1 = 2\omega_0$, shown for comparison by dashed lines in Fig. 15. In this particular case $\omega_m = 3\omega_0/2$ and $\omega_x = \omega_s = \omega_0/2$. With $\psi_0 = n\pi$, equation (4.08) in conjunction with (4.09) gives

$$P(t) = \frac{\hat{e}\omega_0}{2\pi} \cos(\omega_0 t_0/2) \frac{\sin 2\omega_0 t_0 - \sin \omega_0 t_0}{\omega_0 t_0 [1 - (\omega_0 t_0/\pi)^2]}. \quad (5.04)$$

This expression is zero when $\sin 2\omega_0 t_0 - \sin \omega_0 t_0 = 0$, and also when $\cos \omega_0 t_0/2 = 0$. Zero points in the impulse characteristic will occur at uniform intervals $\tau_0 = \pi/\omega_0 = 1/2f_0$. Pulses can accordingly be transmitted at these intervals without mutual interference, or at the same rate as for a low-pass characteristic with the bandwidth to the half-amplitude point equal to ω_0 . This is the same pulse transmission rate as is possible in principle with an ideal band-pass characteristic as shown



(a) AMPLITUDE VS FREQUENCY CHARACTERISTIC



(b) IMPULSE CHARACTERISTIC

Fig. 15 — Particular case of band-pass characteristic with gradual cut-off in which impulse characteristic is zero at intervals $\tau_0 = \frac{1}{2f_0}$.

by the dashed lines in Fig. 15. With a gradual cut-off, however, the phase characteristic will be nearly linear and have a finite slope, so that the above pulse transmission rate can be realized provided $\psi_0 = \pm n\pi$. The same pulse transmission rate can also be attained with vestigial side-band transmission, discussed in section 14.

Another particular case of interest is that shown in Fig. 16, in which $\omega_m = 2\omega_s$. In this case (4.08) becomes with $\psi_0 = \pm n\pi$ and with $\bar{P}(t)$ as

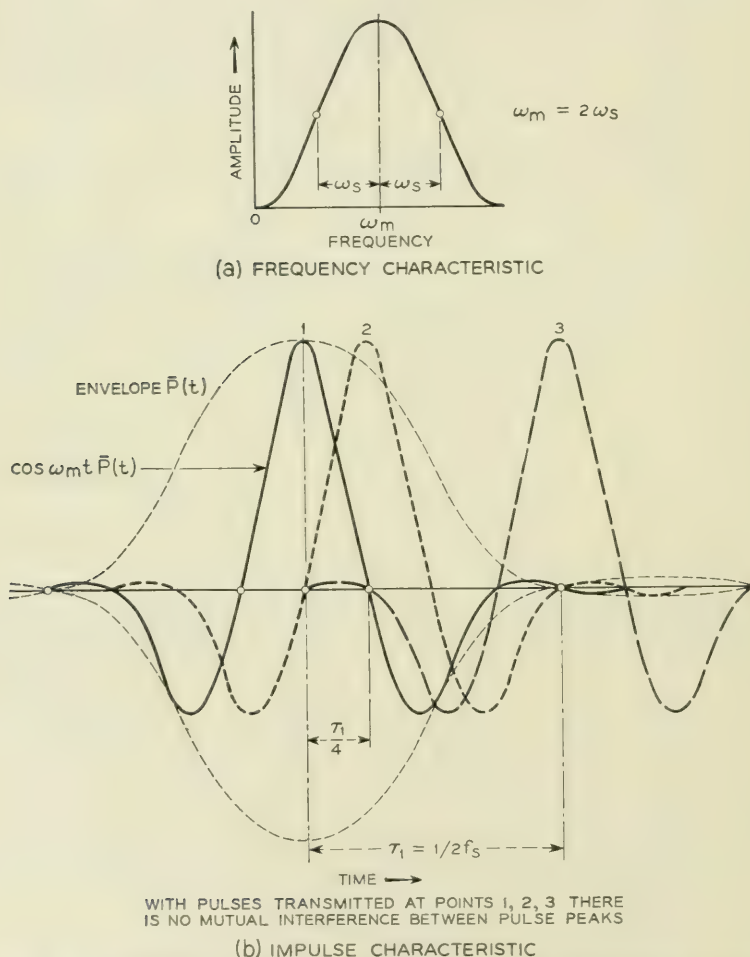


Fig. 16 — Particular case of symmetrical band-pass characteristic for which $\omega_m = 2\omega_s$.

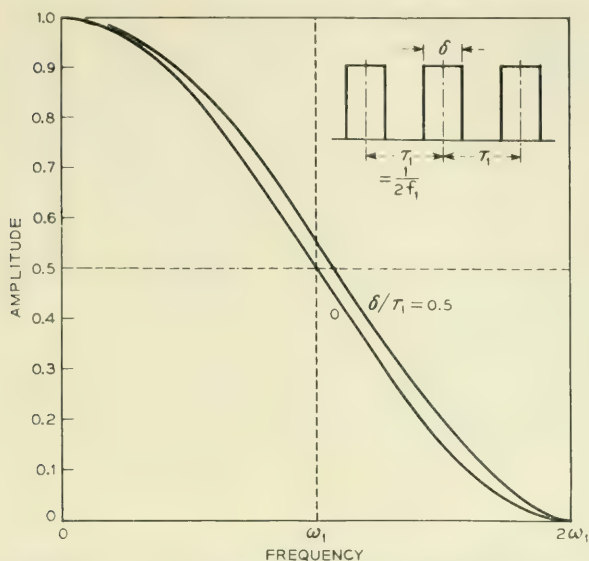


Fig. 17 — Modification of frequency characteristic to obtain same response as for impulses, when pulse duration is prolonged to half the pulse interval.

given by (5.03)

$$\begin{aligned}
 P(t) &= \frac{\delta\omega_m}{\pi} \cos \omega_m t_0 \frac{\sin \omega_m t_0}{\omega_m t_0 [1 - (\omega_m t_0/\pi)^2]}, \\
 &= \frac{\delta\omega_m}{\pi} \frac{\sin 2\omega_m t_0}{2\omega_m t_0 [1 - (\omega_m t_0/\pi)^2]}.
 \end{aligned} \tag{5.05}$$

Pulses can in this case be transmitted without mutual interference between the pulse peaks at the points shown in the above figure. The effective pulse transmission rate is the same as for a low-pass characteristic between $\omega = 0$ and $\omega = 2\omega_m$ with half amplitude at ω_m .*

As mentioned in Section 2, when pulses of finite duration are employed, the same response as for impulses is obtained if the amplitude characteristic is modified by the factor $(\omega\delta/2)/\sin(\omega\delta/2)$. In Fig. 17 is shown the resultant minor modification in the amplitude characteristic (5.01) when the duration of the pulses is equal to half the pulse interval.

The low-pass and band-pass amplitude characteristics considered above can also be regarded as the spectra of pulses applied to a transmission system having a constant amplitude characteristic over the

* W. R. Bennett and C. B. Feldman originally proposed this type of characteristic in an unpublished memorandum, as a means of matching the bandwidth economy of baseband transmission without inclusion of frequencies near zero.

band of the spectra. If the phase characteristic of the system is linear over this band, the received pulses will have the same shape as the impulse characteristics. It should be recognized, however, that there may be appreciable phase distortion within the transmission band or pulse spectrum, if there are amplitude discontinuities beyond the band resulting from a sharp cut-off by filters. Nevertheless, the type of amplitude characteristic or frequency spectrum considered above has decisive advantages from the standpoint of transmission distortion of the pulses, as shown later, since appreciable phase distortion will ordinarily be confined to the edges of the band where the frequency components of the pulse spectrum have low amplitudes.

Another type of amplitude characteristic resembling that shown in Fig. 13 and frequently considered in connection with pulse transmission is a Gaussian characteristic:

$$A(\omega) = e^{-\sigma\omega^2}. \quad (5.06)$$

The corresponding impulse characteristic is

$$P(t) = \frac{\delta}{2(\pi\sigma)^{1/2}} e^{-t_0^2/4\sigma}. \quad (5.07)$$

If it is assumed that the amplitude is reduced to 1 per cent of the peak value after an interval $t_0 = \pi/\omega_1$, corresponding to the first zero point of an ideal impulse characteristic, it is necessary that $t_0^2/4\sigma = 4.6$, or $\sigma = .54/\omega_1^2$. The corresponding amplitude and impulse characteristics are

$$A(\omega) = e^{-0.54(\omega/\omega_1)^2}, \quad (5.08)$$

and

$$P(t) = \frac{\delta\omega_1}{0.83\pi} e^{-0.46(t_0\omega_1)^2}. \quad (5.09)$$

In Fig. 18 a comparison is made of the two frequency characteristics (5.01) and (5.08) considered above, and of the corresponding impulse characteristics (5.02) and (5.09). The comparison shows that for the same pulse transmission rate and with negligible intersymbol interference, a somewhat wider band must be provided with a Gaussian amplitude characteristic. This is a disadvantage, particularly when the band is restricted within prescribed limits by considerations of interference in adjacent transmission bands, as radio pulse systems.

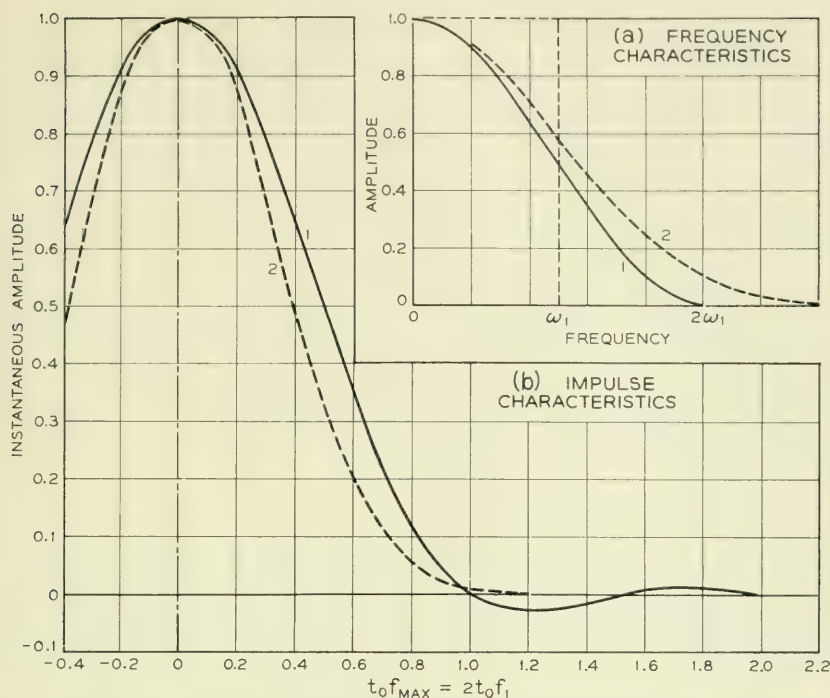


Fig. 18 — Comparison of two representative frequency and impulse transmission characteristics. Frequency characteristic 1: $T(\omega) = \frac{1}{2}[1 + \cos \pi \omega 2\omega_1]$. Frequency characteristic 2: $T(\omega) = \exp - 0.54(\omega/\omega_1)^2$.

Amplitude characteristic 1 of Fig. 18 has certain properties, aside from the linearity of the associated phase characteristic, which makes it preferable to a Gaussian as well as other types of amplitude characteristics for most pulse systems. The corresponding impulse characteristic has zero points at intervals $\tau_1 = 1/2f_1$ with the minimum possible oscillation consistent with this property for a given bandwidth. This permits the use of this impulse characteristic for pulse systems with discrete pulse positions with minimum intersymbol interference and considerable tolerance on synchronization. Since the oscillation in the impulse characteristic is inappreciable, it can also be used for pulse systems without discrete pulse positions and with other methods of detection than synchronized instantaneous sampling. In view of these attributes, an amplitude characteristic of the above type, rather than a constant amplitude characteristic with sharp cut-off, may be regarded as ideal when various physical requirements for practicable pulse systems are taken into consideration.

6. PULSE ECHOES FROM PHASE DISTORTION

For any transmission—frequency characteristic the corresponding impulse characteristics can be determined from the Fourier integral relation (2.01). This, however, may involve the evaluation of complicated integrals, which in general would require numerical integration and would be a rather elaborate procedure. A preferable method of sufficient accuracy in most engineering applications is to employ the theoretical solutions given previously for various ideal transmission characteristics with a linear phase shift as a point of departure or first approximation. A satisfactory second approximation can in many instances be secured by evaluating the transmission distortion resulting from a sinusoidal deviation in the phase characteristic. Furthermore, any type of deviation in the phase characteristic can in principle be represented by a Fourier series in terms of harmonic sinusoidal deviations.

Aside from the circumstance that in many cases a sine deviation in the phase characteristic affords a fairly satisfactory approximation to actual phase distortion it has the advantage in theoretical formulation that it permits determination of the resultant pulse distortion by the method of “paired echoes.” In the usual application of this method only small phase deviations are considered resulting in a single pair of pulse or signal echoes of small amplitude, and the method is then particularly simple.^{5, 6} When delay distortion is appreciable, however, as is frequently the case in wire circuits, it becomes necessary to consider a large number of pulse or signal echoes of considerable amplitude. Since the amplitudes of the pulse echoes may be obtained from available tables of Bessel Functions, the determination of the echoes is, nevertheless, simple in procedure and the determination of the shape of the distorted pulses or other signals not too elaborate.

A given amplitude characteristic within the transmission band may be associated with various phase characteristics, depending on the shape of the amplitude characteristic outside the transmission band and also on whether or not a minimum phase shift system is involved. It is therefore permissible to consider the effect of various departures from a given phase characteristic independent of the amplitude characteristic within the transmission band.

With a sinusoidal departure from a given phase characteristic $\psi_0(\omega)$ as shown in Fig. 19, the modified phase function becomes

$$\psi(\omega) = \psi_0(\omega) - b \sin \omega\tau. \quad (6.01)$$

With

$$T_0(i\omega) = A_0(\omega) e^{-i\tau_0(\omega)}$$

the modified transmission-frequency characteristic becomes

$$T(i\omega) = T_0(i\omega) e^{ib \sin \omega \tau}, \quad (6.02)$$

which, inserted in (2.01) gives

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T_0(i\omega) e^{ib \sin \omega \tau} e^{i\omega t} d\omega. \quad (6.03)$$

The following relation (Jacobi's expansion) in which $J_1, J_2, \dots$ are Bessel Functions in their usual notation can now be employed⁷

$$\begin{aligned} e^{ib \sin \omega \tau} &= J_0(b) + J_1(b)[e^{i\omega \tau} - e^{-i\omega \tau}] \\ &\quad + J_2(b)[e^{2i\omega \tau} + e^{-2i\omega \tau}] \\ &\quad + J_3(b)[e^{3i\omega \tau} - e^{-3i\omega \tau}] \\ &\quad + J_4(b)[e^{4i\omega \tau} + e^{-4i\omega \tau}] + \dots \end{aligned} \quad (6.04)$$

Let $P_0(t)$ designate the shape of the received pulse or other signal for a transmission frequency characteristic $T_0(i\omega)$ obtained from (6.03) with $b = 0$. In view of (6.04) the solution of (6.03) may then be written

$$\begin{aligned} P(t) &= J_0(b)P_0(t) + J_1(b)[P_0(t + \tau) - P_0(t - \tau)] \\ &\quad + J_2(b)[P_0(t + 2\tau) + P_0(t - 2\tau)] \\ &\quad + J_3(b)[P_0(t + 3\tau) - P_0(t - 3\tau)] \\ &\quad + J_4(b)[P_0(t + 4\tau) + P_0(t - 4\tau)] + \dots \end{aligned} \quad (6.05)$$

The shape of the received pulse or signal $P(t)$ is thus obtained by superposing an infinite sequence of pulses or signals of shape $P_0(t)$. The peak amplitudes of the pulse or signal echoes and the times at which they occur with respect to $t = 0$ are given in the following table. The reference point $t = 0$ is arbitrarily selected to coincide with the peak of the pulse $P_0(t)$:

Time	-3τ	-2τ	τ	0	τ	2τ	3τ
Amplitude	$J_3(b)$	$J_2(b)$	$J_1(b)$	$J_0(b)$	$-J_1(b)$	$J_2(b)$	$-J_3(b)$

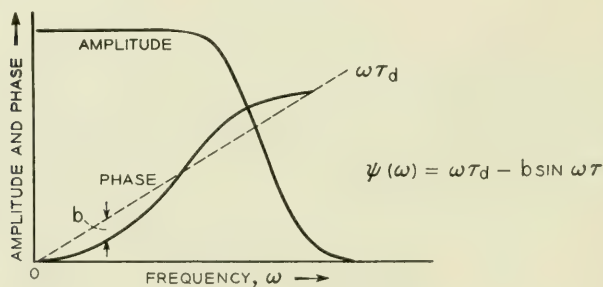
A sufficient number of echoes must be considered until their peak amplitudes become negligible.

The superposition of echoes to obtain the resultant pulse is illustrated in Fig. 20. Instead of plotting the various echoes and combining them

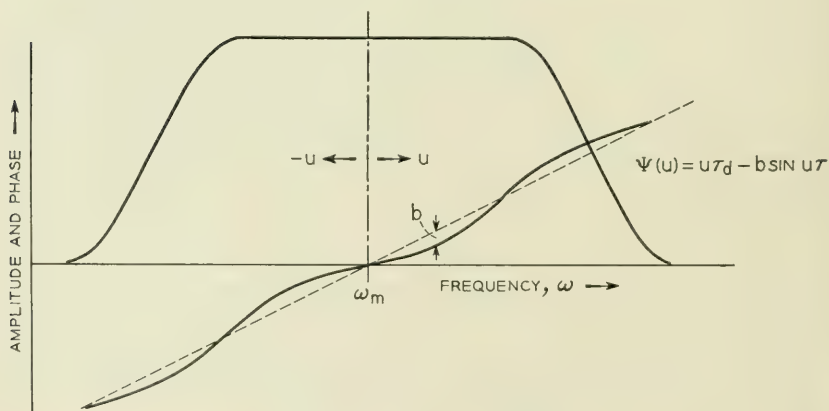
into a resultant pulse or signal as in Fig. 20(c) the equivalent and less laborious method shown in (d) can be employed. With the latter method the pulse P_0 is plotted with reversed time scale and its peak made to coincide with the point for which the amplitude of the resultant pulse P is to be determined. The amplitude of P is determined as indicated in the figure. In the particular case where the original phase characteristic ψ_0 is linear, the pulses $P_0(t)$ will be symmetrical with respect to their peak amplitude, and this assumption will be made in the following applications.

For amplitudes $b \ll 1$, the Bessel Functions become negligible except for J_0 and J_1 , which are given by $J_0(b) \cong 1$ and $J_1(b) \cong b/2$, so that (6.05) becomes

$$P(t) = P_0(t) + \frac{b}{2} P_0(t + \tau) - \frac{b}{2} P_0(t - \tau). \quad (6.06)$$



(a) LOW-PASS CHARACTERISTIC



(b) BANDPASS CHARACTERISTIC

Fig. 19 — Low-pass and band-pass characteristics with sinusoidal phase distortion.

For amplitudes $b > 1$, it is necessary to consider a greater number of echoes, as will be evident from Table I for $b = 1, 2, 5, 10$ and 15 radians.⁸

The preceding equations apply to low-pass characteristics and also to symmetrical bandpass characteristics, as shown in Fig. 19. In the latter case $\alpha(u) = \alpha(-u)$ and $\Psi(-u) = -\Psi(u)$ in (2.10) and (2.11) so that $R_+ = R_-$ and $Q_+ = Q_-$ and (2.09) becomes

$$P(t) = \cos(\omega_r t - \psi_r) R(t), \quad (6.07)$$

where $R(t) = R_+ + R_-$ and $\omega_r = \omega_m =$ midband frequency. The envelope $R(t)$ is accordingly obtained by replacing $P_0(t)$ in (6.05) by $R_0(t)$, the envelope in the absence of phase distortion.

In Fig. 21 is shown a particular case of a sine deviation in the phase characteristic and the corresponding delay distortion, which approximates that encountered in many instances. For a low-pass system the phase and delay distortion would be as shown for $u > 0$. In this particu-

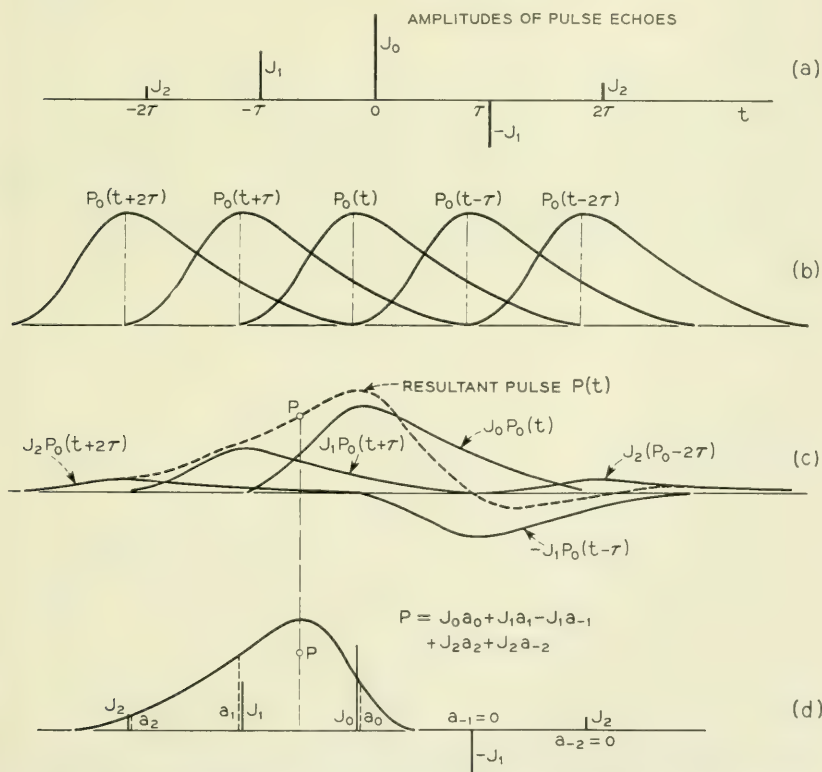


Fig. 20 — Determination of resultant pulse by superposition of pulse echoes.

TABLE I — AMPLITUDES OF ECHOES, $J_n(b)$

	$b = 1$	2	5	10	15
$n = 0$	0.7652	0.2239	-0.1776	-0.2459	-0.0142
1	0.4401	0.5767	-0.3276	0.0434	0.2051
2	0.1149	0.3528	0.0466	0.2546	0.0416
3	0.0196	0.1289	0.3648	0.0584	-0.1940
4		0.0340	0.3912	-0.2196	-0.1192
5			0.2611	-0.2341	0.1305
6			0.1310	-0.0145	0.2061
7			0.0534	0.2167	0.0345
8			0.0184	0.3179	-0.1740
9			0.0055	0.2919	-0.2200
10				0.2075	-0.0901
11				0.1231	0.1000
12				0.0634	0.2367
13				0.0290	0.2787
14				0.0120	0.2464
15				0.0045	0.1813
16					0.1162
17					0.0665
18					0.0346
19					0.0166
20					0.0073

lar case the maximum amplitude b is at the maximum frequency $\omega_{\max} = 2\omega_1$, so that $\sin \omega\tau = 1$, or $\omega\tau = \pi/2$, for $\omega = 2\omega_1$. Hence the interval between pulse echoes is $\tau = \pi/4\omega_1 = 1/8f_1$. The interval τ is accordingly $1/4$ the interval $\tau_1 = 1/2f_1$ required for the pulse $P_0(t)$ to reach zero amplitude in the absence of phase distortion.

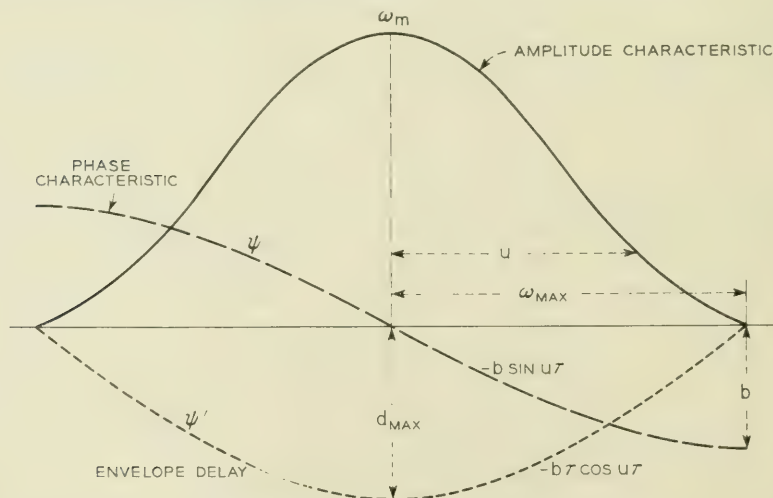


Fig. 21 — Particular case of sinusoidal phase deviation.

For the particular case illustrated in the above figure, the delay distortion is given by

$$d\psi/d\omega = -b\tau \cos \omega\tau.$$

When $\omega = 0$

$$d\psi/d\omega = -b\tau = -d_{\max}. \quad (6.08)$$

When $\omega\tau = \tau\omega_{\max} = \pi/2$

$$d\psi/d\omega = 0.$$

Hence

$$d\psi/d\omega = -d_{\max} \cos (\omega\pi/2\omega_{\max}). \quad (6.09)$$

With $\tau = \pi/2\omega_{\max}$ and $b\tau = d_{\max}$, the following relation is obtained

$$b = \frac{2}{\pi} \omega_{\max} d_{\max} = 4f_{\max} d_{\max}. \quad (6.10)$$

In Fig. 22 are shown the positions of the pulse echoes for the above case on a numerical scale $tf_{\max}$, together with their amplitudes for $b = 5$ radians. On this scale the interval between pulse echoes $\tau = 1/4f_{\max}$ is $1/4$. In the same figure is shown an assumed pulse shape in the absence of phase distortion, which is the same as shown in Fig. 13, except that the small tail has been neglected. The peak of the pulse is taken at $tf_{\max} = -0.75$, and the amplitude of the resultant pulse at the

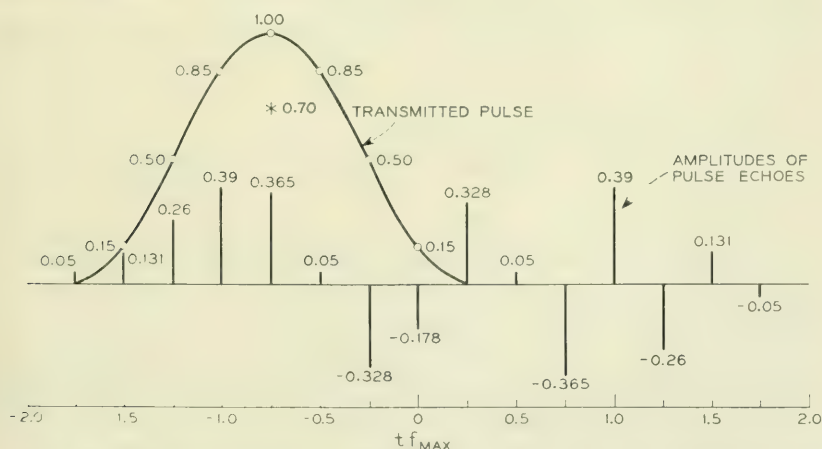


Fig. 22 — Illustrative example of calculation of impulse characteristic shown in Fig. 23, by method illustrated in Fig. 20(d).

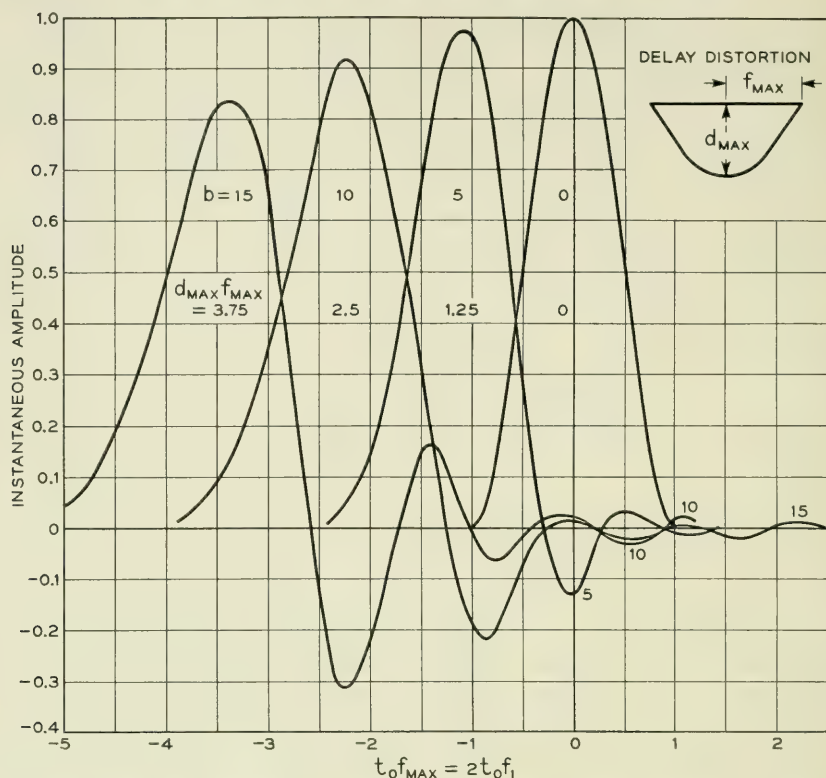


Fig. 23 — Impulse transmission characteristics with cosine variation in delay.

corresponding point obtained by the method illustrated in d of Fig. 20 is

$$\begin{aligned}
 P &= 1 \cdot 0.365 + 0.85 (0.39 + 0.05) \\
 &\quad + 0.5 (0.26 - 0.328) + 0.15 (0.131 - 0.178) \\
 &= 0.70.
 \end{aligned}$$

In Fig. 23 are shown the resultant pulses obtained by the above method for various values of b and the corresponding values of $d_{MAX} f_{MAX}$. Since the interval between pulse echoes is small in relation to the duration of the pulse $P_0(t)$, as seen from Fig. 22, the individual pulse echoes cannot be discerned in the resultant pulses shown in Fig. 23. It will be noticed that as b increases, the pulses are received with decreasing transmission delay, which is due to the choice of reference delay in the delay distortion curve. That is, as d_{MAX} or b is increased, the delay becomes increas-

ingly negative with respect to $d_{\max} = 0$ used for reference. The curves apply to a band-pass system as indicated in the figure, and also to a low-pass system having the delay distortion shown above the midband frequency of the band-pass system.

An improved approximation to phase distortion is sometimes obtained by considering two sine deviations in the phase characteristic.

If the phase characteristic is given by

$$\psi(\omega) = \psi_0(\omega) - b' \sin \omega\tau - b'' \sin \omega\tau', \quad (6.11)$$

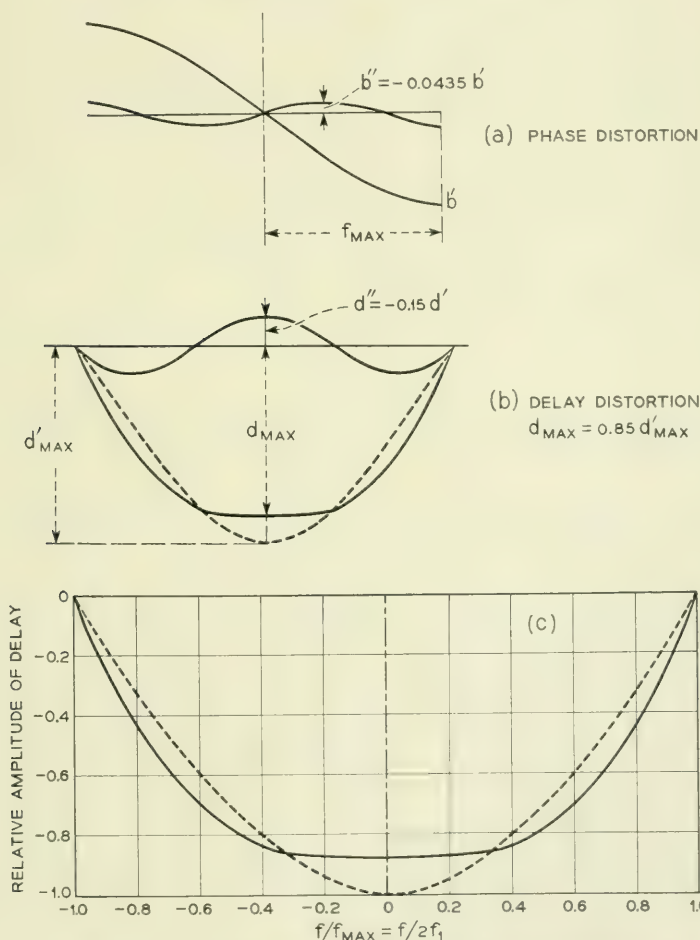


Fig. 24 — Shape of delay distortion with combined fundamental and third harmonic cosine variation in delay.

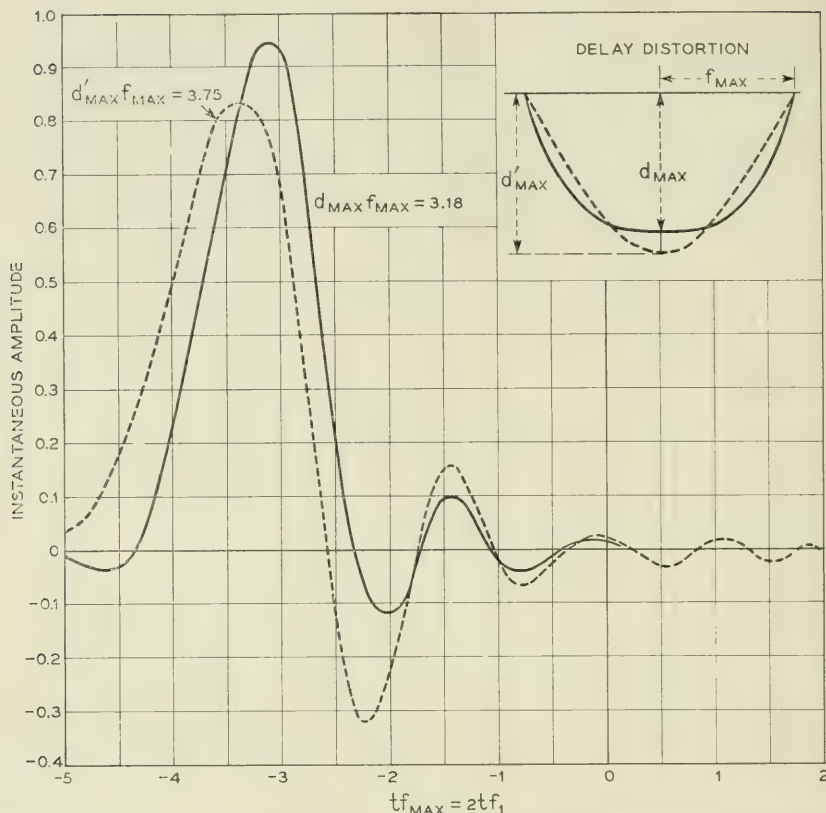


Fig. 25 — Comparison of impulse characteristics with fundamental and combined fundamental and third harmonic cosine variation in delay as in Fig. 24.

the combined effect of the two sine deviations is obtained by first determining the effect of $-b' \sin \omega \tau'$ from (6.05). The value of $P(t) = P_1(t)$ thus obtained from (6.05) with $b = b'$ and $\tau = \tau'$ is next substituted for $P_0(t)$ in (6.05), with $b = b''$ and $\tau = \tau''$ to evaluate the effect of $-b'' \sin \omega \tau''$. That is, the system is considered to consist of a tandem arrangement of two components, the first with a phase distortion $-b' \sin \omega \tau$ and the second with phase distortion $-b'' \sin \omega \tau''$.

In Fig. 24 is shown a particular case in which the second component is a triple harmonic of the first with amplitude $b'' = -0.0435b'$. This results in an improved approximation to the delay distortion encountered in certain wire facilities, where the band is sharply confined by filters. In Fig. 25 is shown the pulse shape for this case with $b' = 15$ radians, together with that for a single sine deviation of $b = 15$ radians.

It will be recognized from the above that as the number of sine com-

ponents required to represent a given phase distortion increases, the determination of the resultant pulse becomes rather laborious, unless the sine deviations are all small in amplitude. In the latter case each sine deviation corresponds in a first approximation to a single pair of echoes, so that the effect of a number of sine deviations can be obtained by direct superposition.

7. PULSE ECHOES FROM AMPLITUDE DISTORTION

Departures from a given amplitude characteristic may in certain cases be approximated by a single cosine variation, as illustrated in Fig. 26. Since the amplitude characteristic is an even function of ω , any departure from a given amplitude characteristic may be represented by a cosine Fourier series. The effect of a cosine variation in the amplitude characteristic is therefore of basic interest.

A cosine variation will in general be accompanied by a change in the phase characteristic, as discussed in Section 1, but it will first be assumed that phase correction is employed to maintain a fixed phase characteristic.

Let $A_0(\omega)$ be the original amplitude characteristic and let the modified amplitude characteristic be of the form

$$A(\omega) = A_0(\omega)[1 + a \cos \omega \tau]. \quad (7.01)$$

Equation (2.01) for the impulse transmission characteristic then becomes, with $T_0(i\omega) = A_0(\omega)e^{-i\psi_0(\omega)}$,

$$\begin{aligned} P(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} T_0(i\omega) \left[1 + \frac{a}{2} (e^{i\omega\tau} + e^{-i\omega\tau}) \right] e^{i\omega t} d\omega, \\ &= P_0(t) + \frac{a}{2} P_0(t + \tau) + \frac{a}{2} P_0(t - \tau). \end{aligned} \quad (7.02)$$

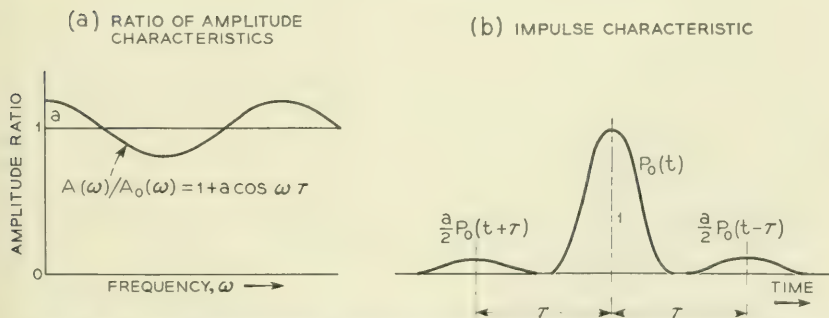


Fig. 26 — Pulse echoes from cosine variation in amplitude characteristic with out change in phase characteristic.

There will thus be pulse or signal echoes of amplitude $a/2$ at the time τ before and after the main pulse as illustrated in Fig. 26.

With a cosine variation in the attenuation rather than in the amplitude characteristic, the modified amplitude characteristic becomes

$$A(\omega) = A_0(\omega) e^{a \cos \omega \tau}, \quad (7.03)$$

and the modified impulse characteristic

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} T_0(i\omega) e^{a \cos \omega \tau} e^{i\omega t} d\omega. \quad (7.04)$$

The expansion corresponding to (6.04) is in this case:

$$\begin{aligned} e^{a \cos \omega \tau} = & I_0(a) + I_1(a)(e^{i\omega \tau} + e^{-i\omega \tau}) \\ & + I_2(a)(e^{2i\omega \tau} + e^{-2i\omega \tau}) \\ & + I_3(a)(e^{3i\omega \tau} + e^{-3i\omega \tau}), + \dots \end{aligned} \quad (7.05)$$

where $I_1, I_2 \dots$ are Bessel functions for imaginary arguments in their usual notation.

The resultant modified impulse characteristic in this case becomes

$$\begin{aligned} P(t) = & I_0(a)P_0(t) + I_1(a)[P_0(t + \tau) + P_0(t - \tau)] \\ & + I_2(a)[P_0(t + 2\tau) + P_0(t - 2\tau)] \\ & + I_3(a)[P_0(t + 3\tau) + P_0(t - 3\tau)] + \dots \end{aligned} \quad (7.06)$$

which can be interpreted in a similar way as discussed for (6.05). For small values of a , $I_0(a) \cong 1$, $I_1(a) \cong a/2$ and the remaining terms in (7.06) negligible, so that (7.02) is obtained.

As discussed in Section 1, when the amplitude characteristic is modified in accordance with (7.01), the resultant modification in the phase characteristic is in accordance with (1.13)

$$\psi_1 = 2 \tan^{-1} \frac{r \sin \omega \tau}{1 + r^2 \cos \omega \tau}. \quad (7.07)$$

The modified transmission-frequency characteristic is in this case

$$T(i\omega) = T_0(i\omega)(1 + a \cos \omega \tau)e^{-i\psi_1}, \quad (7.08)$$

which can be transformed into

$$\begin{aligned} T(i\omega) = & T_0(i\omega) \frac{1}{1 + r^2} (1 + re^{-i\omega \tau})^2, \\ = & T_0(i\omega) \frac{1}{1 + r^2} (1 + 2re^{-i\omega \tau} + r^2 e^{-2i\omega \tau}). \end{aligned} \quad (7.09)$$

Thus, with a cosine variation in the amplitude characteristic in accordance with (7.01), accompanied by a minimum phase shift change in the phase characteristic in accordance with (7.07), the modified impulse characteristic becomes

$$P(t) = \frac{1}{1+r^2} [P_0(t) + 2rP_0(t-\tau) + r^2P_0(t-2\tau)], \quad (7.10)$$

where

$$r = \frac{1}{a} [1 - \sqrt{1-a^2}]. \quad (7.11)$$

The received pulse or signal $P(t)$ will thus consist of three components each having the same shape as the pulse or signal $P_0(t)$, but differing in amplitude and displaced in time, as indicated in Fig. 27.

For small values of the amplitude a of the cosine deviation, $r \cong a/2$ and $1+r^2 \cong 1$, so that

$$P(t) = P_0(t) + aP_0(t-\tau) + \frac{a^2}{4}P_0(t-2\tau). \quad (7.11)$$

The solution for a somewhat similar case given elsewhere,⁹ has an infinite number of echoes, with the second echo given by $a^2P_0(t-2\tau)$ rather than $(a^2/4)P_0(t-2\tau)$ as above. In the case referred to, the amplitude deviation is in a first approximation $a \cos \omega\tau$, but there are additional terms in $\cos 2\omega\tau$, $\cos 3\omega\tau$ etc, which are responsible for the different amplitude of the second echo and for the infinite sequence of echoes.

With a cosine modification in the attenuation characteristic as given by (7.03), there will be a corresponding sine modification in the phase characteristic in accordance with (1.11). The modified transmission-

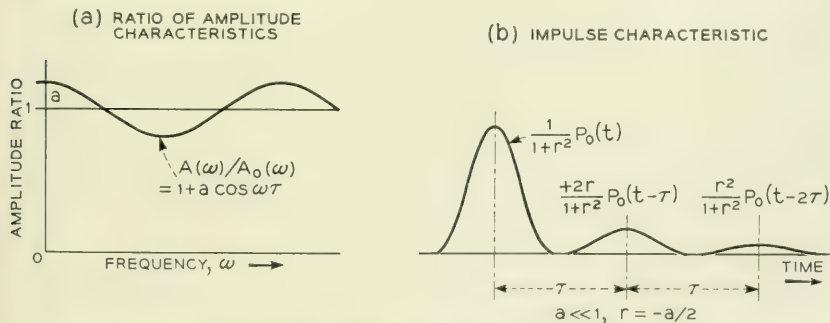


Fig. 27 — Pulse echoes from cosine variation in amplitude characteristic with associated minimum phase shift variation in phase characteristic.

frequency characteristic is in this case

$$\begin{aligned} T(i\omega) &= T_0(i\omega)e^{a(\cos \omega\tau - i \sin \omega\tau)}, \\ &= T_0(i\omega)e^{ae^{-i\omega\tau}}, \\ &= T_0(i\omega) \left[1 + ae^{-i\omega\tau} + \frac{a^2}{2!} e^{-2i\omega\tau} + \frac{a^3}{3!} e^{-3i\omega\tau} + \dots \right]. \end{aligned} \quad (7.12)$$

The modified impulse characteristics is in this case

$$\begin{aligned} P(t) &= P_0(t) + aP_0(t - \tau) + \frac{a^2}{2!} P_0(t - 2\tau) \\ &\quad + \frac{a^3}{3!} P_0(t - 3\tau) + \dots \end{aligned} \quad (7.13)$$

For small values of a both (7.11) and (7.13) give for the modification in the impulse characteristic resulting from a small cosine deviation in the amplitude or attenuation characteristics accompanied by changes in the phase characteristic:

$$P(t) = P_0(t) + a P_0(t - \tau). \quad (7.14)$$

In certain applications it is convenient to regard $P_0(t)$ as a pulse or signal applied to a transmission line and $P(t)$ as the received pulse or signal with a cosine deviation in the amplitude characteristic of the transmission line.

In the lower part of Fig. 28 is shown the modification in the received pulses resulting from a slow pronounced cosine deviation in the amplitude characteristic shown at the top. In Fig. 29 is shown the effect of positive and negative cosine variations when the amplitude at zero frequency is held constant, a condition which may be approximated in wire systems as a result of variation in attenuation over the transmission band with temperature. Curve 1 would correspond to a 3.5 db smaller loss at the maximum frequency $2\omega_1$ than at zero frequency, and curve 2 to a 6 db greater loss at the maximum frequency. It will be noticed that pulse distortion as well as the variation in the peak amplitude of the pulses is greater under the first condition, i.e. curve 1. Pulse overlaps can in both cases be avoided by a moderate increase in pulse spacing, and in the first case can be substantially reduced also by a decrease in pulse spacing.

8. FINE STRUCTURE IMPERFECTIONS IN TRANSMISSION CHARACTERISTICS

As a result of imperfections in the transmission medium and in equalization there may be fine structure departures from a nominal transmission characteristic, as illustrated in Fig. 30. They are often caused by

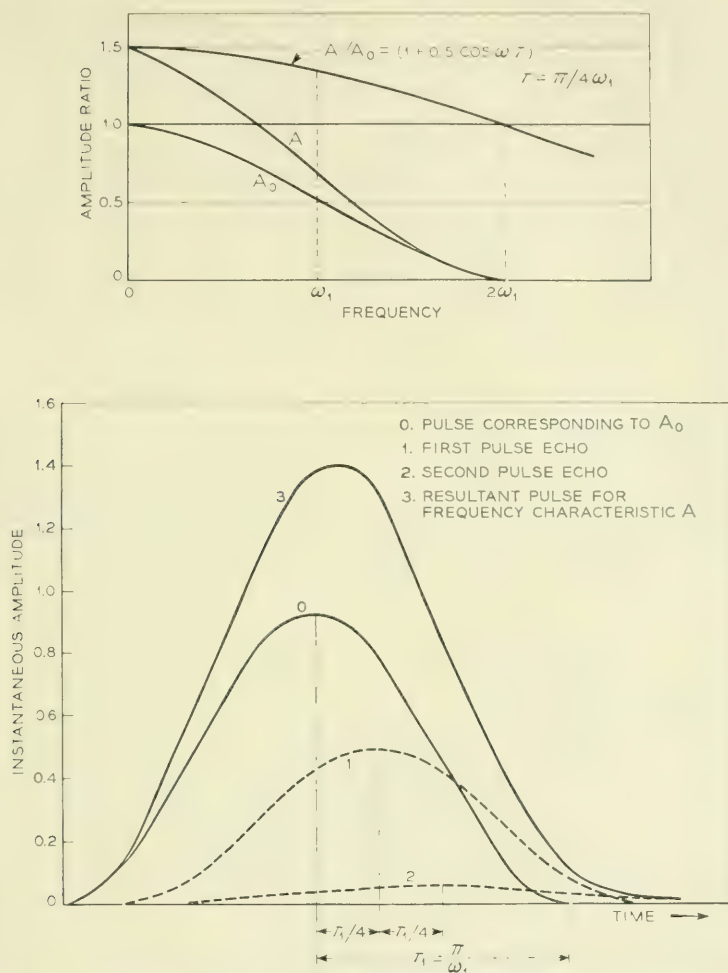


Fig. 28 — Modification of impulse characteristic by slow cosine variation in amplitude characteristic.

echoes in very long lines resulting from impedance mismatches. Fine structure deviations from a specified amplitude characteristic may in principle be represented by a cosine Fourier series, since the amplitude function is an even function of ω . Thus, if the specified amplitude characteristic is $A_0(\omega)$, the actual amplitude characteristic $A(\omega)$ may be represented by an infinite cosine Fourier series as:

$$A(\omega) = A_0(\omega)[1 + a_1 \cos \omega \tau + a_2 \cos 2\omega \tau + \cdots + a_m \cos m\omega \tau + \cdots]. \quad (8.01)$$

The coefficients $a_1, a_2 \cdots a_m \cdots$ are determined in the usual manner by Fourier series analysis to represent the function

$$f(\omega) = \frac{A(\omega)}{A_0(\omega)} = 1 + \alpha(\omega) \quad (8.02)$$

over the frequency band. If $A_0(\omega)$ closely approaches $A(\omega)$ the fine

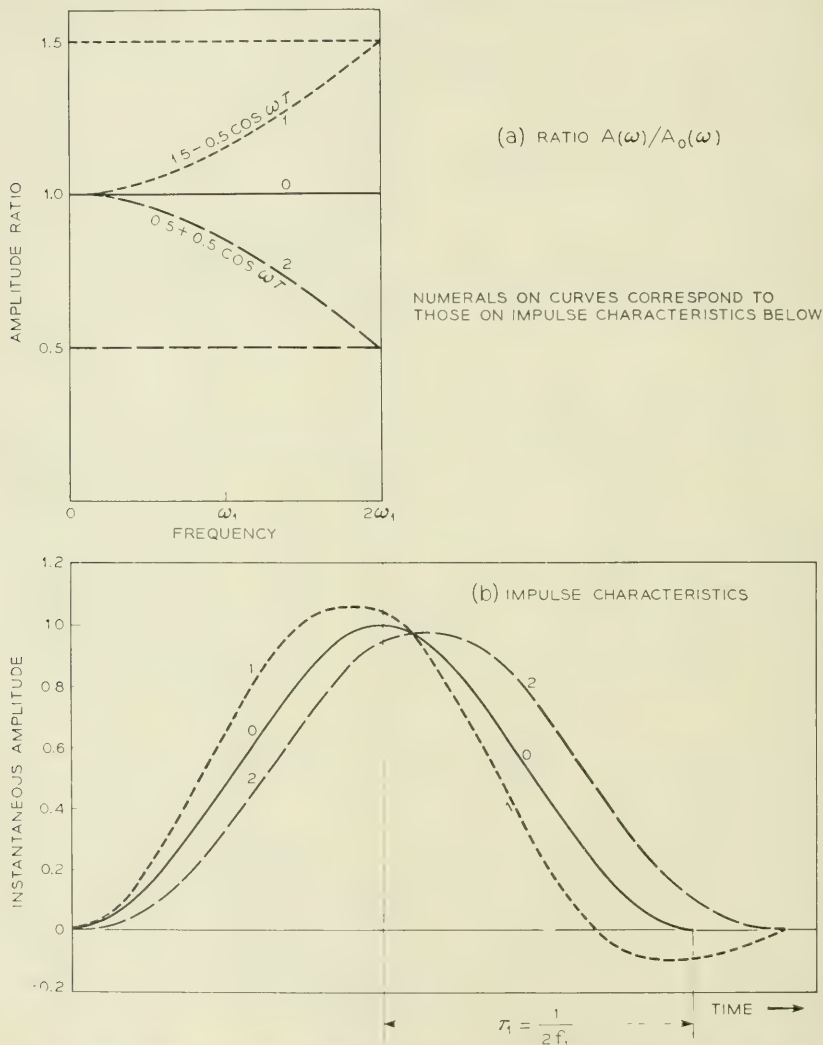


Fig. 29 — Effect of slow cosine variation in amplitude characteristic when amplitude at zero frequency is held constant.

structure departures $\alpha(\omega)$ in the transmission characteristic and hence the coefficients $a_1, a_2 \cdots a_m \cdots$ will be small.

In the above representation $A_0(\omega)$ can also be regarded as the amplitude characteristic of a terminal network or as the frequency spectrum of a pulse applied to a transmission system with an amplitude characteristic $f(\omega) = 1 + \alpha(\omega)$.

In a Fourier series analysis of the deviation in the amplitude characteristic, the fundamental period of the amplitude variation would be selected so that there is one complete cycle between $-\omega_1$ and ω_1 , the cutoff frequency, in which case $\omega_1\tau = \pi$ or

$$\tau = \frac{\pi}{\omega_1}.$$

This is the interval between pulse echoes when the amplitude characteristic is represented by (8.01). It is identical with the interval τ_1 given by (3.02) at which pulses can be transmitted without mutual interference with a constant amplitude transmission frequency characteristic.

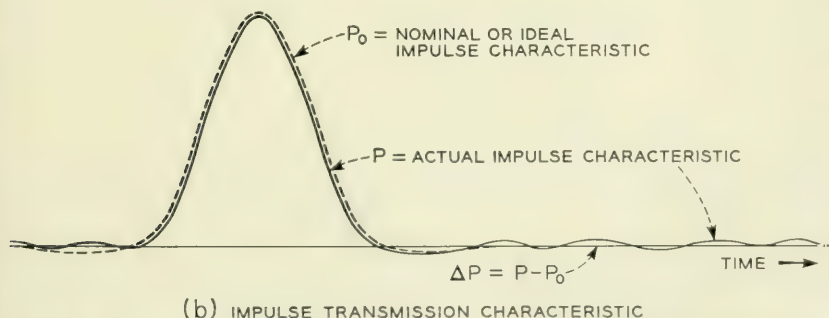
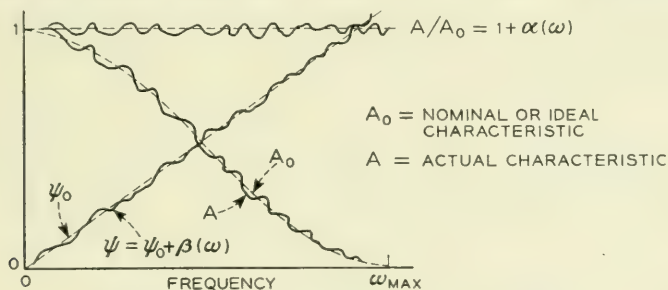
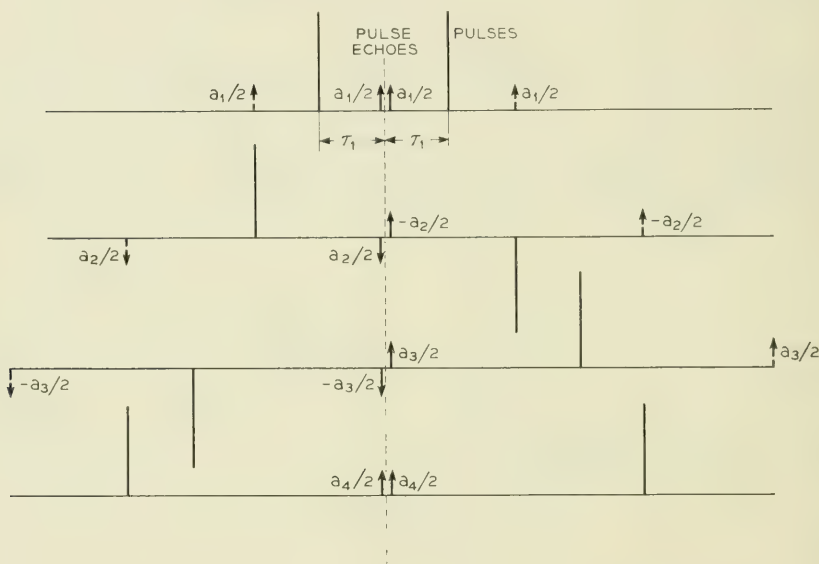


Fig. 30 — Fine structure imperfections in transmission frequency characteristic and resultant prolongation of impulse characteristic.

Assume that pulses of unit peak amplitude but varying polarity are transmitted at intervals $\tau = \tau_1$ and consider the interference with a given pulse from all pulses. As illustrated in Fig. 31, the first preceding and following pulses will in accordance with (7.02) give rise to a pulse echo $\pm a_1/2$ and the second preceding and following pulses to a pulse echo $\pm a_2/2$ etc., where the signs of the echoes depend on the polarity of the pulses and on the signs of the coefficients a_1, a_2 . The resultant intersymbol interference $U_a(t)$ will depend on the polarity of the various

PULSE ECHOES FROM INDIVIDUAL PULSES



RESULTANT PULSE TRAIN AND INTERSYMBOL INTERFERENCE



$$U_a = \frac{a_1}{2} + \frac{a_1}{2} + \frac{a_2}{2} - \frac{a_2}{2} - \frac{a_3}{2} + \frac{a_3}{2} + \frac{a_4}{2} + \frac{a_4}{2} = a_1 + a_4$$

Fig. 31 — Combination of pulse echoes into intersymbol interference for a particular case.

pulses and will thus vary with time. It can have any value assumed by the expression

$$U_a(t) = \pm \frac{a_1}{2} \pm \frac{a_1}{2} \pm \frac{a_2}{2} \pm \frac{a_2}{2} \cdots \pm \frac{a_m}{2} \pm \frac{a_m}{2} + \cdots \quad (8.03)$$

The maximum possible intersymbol interference will thus be the sum of the absolute values of the coefficients a_m .

$$\hat{U}_a = |a_1| + |a_2| + |a_3| + \cdots + |a_m| + \cdots \quad (8.04)$$

In certain pulse systems, such as PAM time division systems, rms intersymbol interference is of main importance, while in others, such as PCM or telegraph systems, peak intersymbol interference is of principal interest. If the fine structure imperfections are regarded as of random nature, in the sense that they are not predictable and vary between systems having the same nominal transmission characteristics, peak intersymbol interference can be estimated from rms interference by applying a peak factor of about 4. With random variation in the amplitude of intersymbol interference, the probability of exceeding 4 times the rms value is in accordance with the normal law about 5×10^{-5} . Peaks in excess of 4 times the rms value will thus be so rare that they can for practical purposes be neglected.

The rms intersymbol interference is equal to the root mean square of all the different values which can be assumed by expression (8.03). This turns out to be equal to the root sum square of the amplitudes $a_m/2$ and $-a_m/2$ of the pulse echoes, or

$$\underline{U}_a = \left[\sum_1^\infty \left(\frac{a_m}{2} \right)^2 + \sum_1^\infty \left(\frac{-a_m}{2} \right)^2 \right]^{1/2} = \left(\frac{1}{2} \sum_1^\infty a_m^2 \right)^{1/2}. \quad (8.05)$$

When a_m are the various coefficients in the Fourier representation of $\alpha(\omega)$ over the frequency band from $-\omega_1$ to ω_1 , the following relation holds.

$$\frac{1}{2} \sum_1^\infty a_m^2 = \frac{1}{2\omega_1} \int_{-\omega_1}^{\omega_1} \alpha^2(\omega) d\omega = \frac{1}{\omega_1} \int_0^{\omega_1} \alpha^2(\omega) d\omega \quad (8.06)$$

where $\alpha(\omega)$ in the present case is given by (8.02) and represents the departure in the ratio $A(\omega)/A_0(\omega)$ from unity.

With (8.06) in (8.05) the following expression is obtained for rms intersymbol interference due to amplitude deviations $\alpha(\omega)$ not accompanied by phase deviations

$$\underline{U}_a = \underline{a} \quad (8.07)$$

where $\underline{a}$ is the rms deviation in $\alpha(\omega)$ over the transmission band as given by

$$\begin{aligned}\underline{a} &= \left[\frac{1}{\omega_1} \int_0^{\omega_1} \alpha^2(\omega) d\omega \right]^{1/2} \\ &= \left[\frac{1}{f_1} \int_0^{f_1} \alpha^2(f) df \right]^{1/2}\end{aligned}\quad (8.08)$$

The rms amplitude deviation expressed in db is

$$\begin{aligned}\underline{a}^{\text{db}} &= 20 \log_{10}(1 + \underline{a}) \\ &\cong 8.69 \underline{a} \quad \text{when } \underline{a} < 0.1\end{aligned}\quad (8.09)$$

A corresponding analysis can be made for fine structure imperfections in the phase characteristic. The deviation $\beta(\omega) = \psi(\omega) - \psi_0(\omega)$ from a prescribed phase characteristic $\psi_0(\omega)$ may in this case be represented by a sine Fourier series since the phase characteristic is an odd function of ω :

$$\beta(\omega) = b_1 \sin \omega\tau + b_2 \sin 2\omega\tau + \cdots + b_m \sin m\omega\tau + \cdots \quad (8.10)$$

The resultant peak intersymbol interference becomes

$$\hat{U}_b = |b_1| + |b_2| + \cdots + |b_m| + \cdots \quad (8.11)$$

and the rms intersymbol interference

$$\underline{U}_b = \left[\frac{1}{2} \sum_1^\infty b_m^2 \right]^{1/2} = \underline{b}, \quad (8.12)$$

where $\underline{b}$ is the rms phase deviation in radians as given by

$$\begin{aligned}\underline{b} &= \left[\frac{1}{\omega_1} \int_0^{\omega_1} \beta^2(\omega) d\omega \right]^{1/2}, \\ &= \left[\frac{1}{f_1} \int_0^{f_1} \beta^2(f) df \right]^{1/2}.\end{aligned}\quad (8.13)$$

In the above derivation, the amplitude and phase deviations were assumed independent of each other. The resultant rms intersymbol interference from both is in this case

$$\underline{U} = (\underline{U}_a^2 + \underline{U}_b^2)^{1/2} = (\underline{a}^2 + \underline{b}^2)^{1/2}. \quad (8.14)$$

This relationship, applying to an ideal transmission characteristic, has been established by a different method in a paper by W. R. Bennett.¹⁰

From (7.05) it will be seen that with minimum phase shift relationships a small cosine deviation of amplitude a_m in the amplitude characteristic will be accompanied by a phase deviation $b_m = a_m$. Hence in this case (8.14) gives

$$\underline{U} = 2^{1/2} \underline{a} \quad (8.15)$$

This also follows when it is considered that in this case all the pulse echoes occur after the main pulse, and have amplitudes $a_1, a_2 \dots a_m$. The root sum square of the amplitudes is in this case $[\sum_1^{\infty} a_m^2]^{1/2}$, which is greater than $\underline{U}_a$ as given by (8.05) by the factor $2^{1/2}$.

The above analysis was based on an infinite sequence of pulse echoes, which combine to give the proper pulse distortion but may be regarded as fictitious in nature. The assumption of an infinite sequence of pulse echoes can be avoided by a different method of analysis outlined below, which does not involve the assumption that the coefficients are known from a Fourier series analysis, and furthermore, does not assume an ideal amplitude characteristic with a sharp cut-off as above.

Let $Ae^{-i\psi}$ and $A_0e^{-i\psi_0}$ designate two transmission frequency characteristics, where A, A_0, ψ and ψ_0 are functions of ω , which for convenience is omitted in the following. The squared absolute value of the difference in the transmission frequency characteristics is then

$$|Ae^{-i\psi} - A_0e^{-i\psi_0}|^2 = A_0^2[2(1 - \cos \beta)(1 + \alpha) + \alpha^2], \quad (8.16)$$

where $\alpha = \alpha(\omega) = (A - A_0)/A_0$ represents the deviation in the ratio of the amplitude characteristics from unity and $\beta = \beta(\omega) = \psi - \psi_0$ the deviation in the phase characteristic.

Let P and P_0 designate the impulse characteristics corresponding to the above transmission frequency characteristics, and let $\Delta P = P - P_0$. Assume that unit impulses of varying polarity are transmitted at uniform intervals τ_1 . The rms value of ΔP over the interval τ_1 in relation to the maximum amplitude $P(0)$ of the received pulses, or the rms inter-symbol interference $\underline{U}$, is then given by

$$\begin{aligned} \underline{U} &= \frac{1}{P(0)} \left[\frac{1}{\tau_1} \int_{-\infty}^{\infty} (\Delta P)^2 dt \right]^{1/2}, \\ &= \frac{1}{P(0)} \left[\frac{1}{\pi \tau_1} \int_0^{\infty} A^2 [2(1 - \cos \beta)(1 + \alpha) + \alpha^2] d\omega \right]^{1/2}. \end{aligned} \quad (8.17)$$

For small values of α and β , this expression becomes

$$\underline{U} = \frac{1}{P(0)} \left[\frac{1}{\pi \tau_1} \int_0^{\infty} A_0^2 (\alpha^2 + \beta^2) d\omega \right]^{1/2}. \quad (8.18)$$

If α and β are random variables representing fine structure deviations uniformly distributed over the transmission band, it is permissible to simplify (8.18) to:

$$\underline{U} = \eta \left(\frac{1}{\omega_1 \tau_1} \right)^{1/2} (\underline{a}^2 + \underline{b}^2)^{1/2}, \quad (8.19)$$

where

$$\eta = \frac{1}{\pi P(0)} \left(\omega_1 \int_0^{\omega_{\max}} A_0^2 d\omega \right)^{1/2}, \quad (8.20)$$

$$\underline{a} = \left(\frac{1}{\omega_{\max}} \int_0^{\omega_{\max}} \alpha^2 d\omega \right)^{1/2}, \quad \text{and} \quad (8.21)$$

$$\underline{b} = \left(\frac{1}{\omega_{\max}} \int_0^{\omega_{\max}} \beta^2 d\omega \right)^{1/2}, \quad (8.22)$$

where $\omega_{\max}$ is defined as in Fig. 30 and ω_1 is the bandwidth at the half amplitude point.

For a transmission characteristic with linear phase shift, aside from small random imperfections as considered here:

$$P(0) = \frac{1}{\pi} \int_0^{\omega_{\max}} A_0 d\omega. \quad (8.23)$$

For the particular case of a transmission characteristic with constant amplitude between $\omega = 0$ and $\omega_1 = \omega_{\max}$, $\eta = 1$. Pulses would in this case be transmitted at intervals $\tau_1 = \pi/\omega_1$ so that $\pi/\omega_1 \tau_1 = 1$ and (8.19) is identical with (8.14).

For a transmission characteristic of the type shown in Fig. 13, pulses would also be transmitted at intervals $\tau_1 = \pi/\omega_1$ so that $\pi/\omega_1 \tau_1 = 1$. In this case $\omega_{\max} = 2\omega_1$, and evaluation of (8.20) gives $\eta = 3^{1/2}/2 = 0.866$. Rms intersymbol interference is thus reduced by the factor 0.866, for the same values of $\underline{a}$ and $\underline{b}$. However, these are now the rms deviations taken over a band which is twice as great as with a sharp cut-off at ω_1 .

Expressions (8.14) and (8.19) can also be applied to localized imperfections in the amplitude and phase characteristics confined to a narrow portion of the transmission band. This follows when it is considered that such deviations can be represented by Fourier series containing a large number of coefficients, so that the resultant intersymbol interference can attain a great number of different values depending on the sequence of transmitted pulses. A particular case of a localized imperfection in the amplitude characteristic in the form of a low-frequency cut-off is considered in the following section.

9. TRANSMISSION DISTORTION BY LOW FREQUENCY CUT-OFF

A low-frequency cut-off in the transmission frequency characteristic of wire systems is unavoidable with transformers as employed for increased transmission efficiency or other reasons. In single sideband frequency division systems, there is a low-frequency cut-off in individual channels caused by elimination of the carrier and part of the desired sideband. The effect of a low-frequency cut-off can be avoided by employing a symmetrical band-pass characteristic as illustrated in Fig. 16, or more generally by double sideband transmission with a two-fold increase in bandwidth as compared to a low-pass system. It can also be overcome by vestigial sideband transmission with inappreciable bandwidth penalty, but with complications in terminal instrumentation. The effect of a low-frequency cut-off can, furthermore, be reduced without frequency translation as involved in double or vestigial sideband transmission, by certain methods of shaping or transmission of pulses, as discussed in the following, and by certain methods of compensation at the receiving end or at points of pulse regeneration not considered here.

The nature of the pulse distortion resulting from a low-frequency cut-off is illustrated in Fig. 32. If the phase characteristic is assumed linear, the amplitude characteristic may be regarded as made up of two components, in accordance with the following identity:

$$A(\omega) = A_0(\omega) + [A(\omega) - A_0(\omega)], \quad (9.01)$$

where $A_0(\omega)$ is the amplitude characteristic without a low-frequency cut-off and $[A(\omega) - A_0(\omega)]$ a supplementary characteristic of negative amplitude, as indicated in Fig. 32.

The impulse characteristic may correspondingly be written

$$P(t) = P_0(t) + [P(t) - P_0(t)]. \quad (9.02)$$

If the cut-off is confined to rather low frequencies, the impulse characteristic $\Delta P(t) = P(t) - P_0(t)$ will extend over time intervals substantially longer than the duration of $P_0(t)$ or the interval at which pulses are transmitted. The total area under the resultant pulse is always zero.

When a sufficiently long sequence of pulses of one polarity is transmitted, the cumulative effect of the pulse overlaps resulting from the modification $P(t) - P_0(t)$ in the impulse characteristic will be a displacement of the received pulse train, as illustrated in Fig. 33 for various intervals between the pulses. This apparent displacement of the zero line, often referred to as "zero wander," will reduce the margin for dis-

inction between the presence and absence of pulses in a random pulse train. In the particular case when pulses are transmitted at the minimum interval $\tau_1 = 1/2f_1$ possible without intersymbol interference in the absence of a low-frequency cut-off, the pulse train will ultimately vanish when an infinite sequence of pulses of one polarity is transmitted, as illustrated for the last case in Fig. 33.

The number of pulses of one polarity, or nearly all of the same polarity, which can be transmitted before the limiting condition illustrated in Fig. 33 is approached depends on the extent of the low-frequency cut-off. If the low-frequency cut-off is inappreciable, this number may be sufficiently great so that the probability of encountering such a sequence in a random pulse train and resultant errors in reception may be so small that it can be disregarded. The requirement of the low-frequency cut-off which is necessary to this end is evaluated below for pulses transmitted at intervals $\tau_1 = 1/2f_1$.

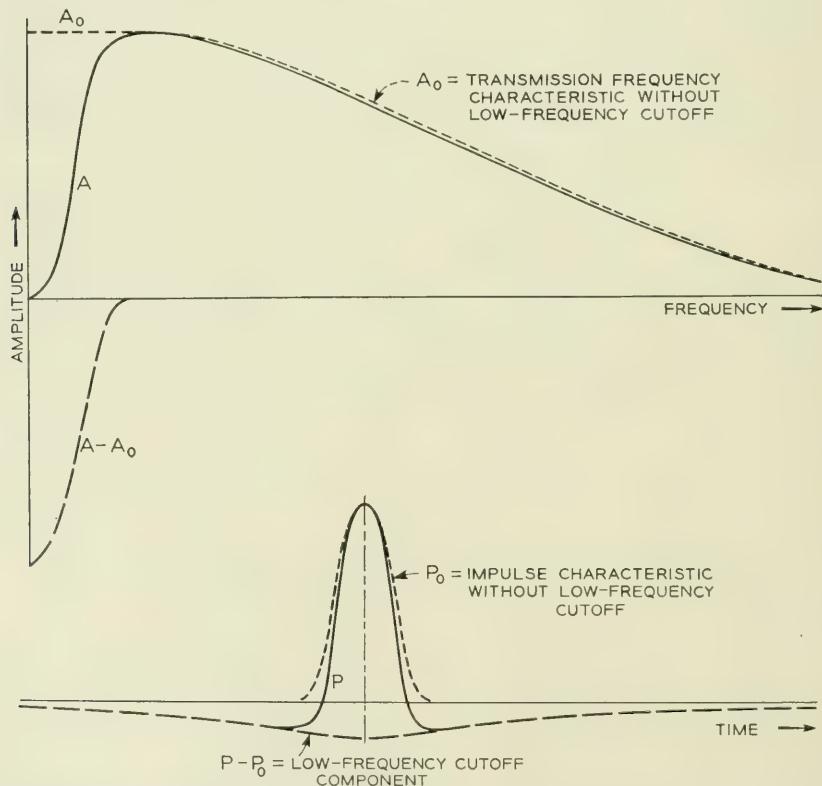


Fig. 32 — Separation of low-frequency cut-off components $A - A_0$ and $P - P_0$ in transmission frequency and impulse characteristics.

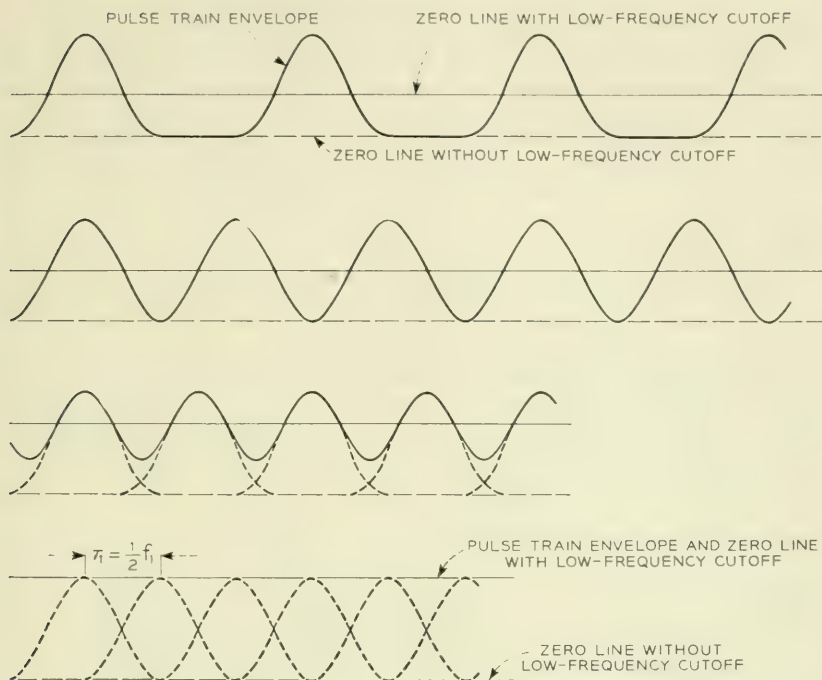


Fig. 33 — Effect of low-frequency cut-off on recurrent pulses as pulse interval is decreased.

If it is assumed that positive and negative impulses are applied at random to the transmission systems at intervals τ_1 , the rms intersymbol interference resulting from a low-frequency cut-off can be evaluated by essentially the same method as employed in Section 8 for fine structure imperfections in the transmission characteristic, provided ω_0 is much smaller than ω_1 . On this basis, rms intersymbol interference in relation to the peak amplitude $P_0(0)$ of the pulses in the absence of a low-frequency cut-off becomes:

$$\underline{U} = \frac{1}{P_0(0)} \left(\frac{1}{\tau_1} \int_{-\infty}^{\infty} [P(t) - P_0(t)]^2 dt \right)^{1/2}, \quad (9.03)$$

$$= \frac{1}{P_0(0)} \left(\frac{1}{\pi \tau_1} \int_0^{\infty} [A(\omega) - A_0(\omega)]^2 d\omega \right)^{1/2}. \quad (9.04)$$

For a transmission characteristic with linear phase shift

$$P_0(0) = \frac{1}{\pi} \int_0^{\infty} A_0(\omega) d\omega. \quad (9.05)$$

For the particular case of sharp cut-offs at ω_0 and ω_1

$$\begin{aligned} A_0(\omega) &= 1 & 0 < \omega < \omega_1, \\ A(\omega) - A_0(\omega) &= -1 & 0 < \omega < \omega_0, \text{ and} \\ P_0(0) &= \omega_1/\pi & \tau_1 = \pi/\omega_1, \end{aligned}$$

and

$$\underline{U} = \frac{\pi}{\omega_1} \left(\frac{\omega_1 \omega_0}{\pi^2} \right)^{1/2} = \left(\frac{\omega_0}{\omega_1} \right)^{1/2}. \quad (9.06)$$

It will be noticed that the same result is obtained from (8.07) with the amplitude deviation $\alpha = [A(\omega) - A_0(\omega)] = -1$ between 0 and ω_0 .

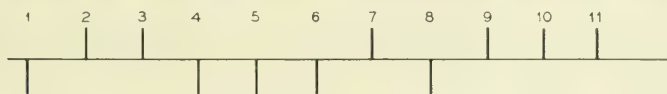
In actual systems, the low-frequency cut-off will be gradual between $\omega = 0$ and ω_0 , rather than abrupt as assumed above. With a linear variation in the amplitude characteristic between 0 and ω_0 , $A(\omega) - A_0(\omega) = (-1 + \omega/\omega_0)A_0(0)$ and $\underline{U} = (\omega_0/3\omega_1)^{1/2}$.

If a sufficient number of pulses of one polarity is transmitted in succession at intervals $\tau_1 = 1/2f_1$ the received pulses will as noted before in the limit be reduced to zero amplitude by the low-frequency cut-off. The maximum pulse distortion resulting from pulse overlaps when a train of pulses as transmitted is thus equal and opposite to the amplitude $P_0(0)$ of the received pulses in the absence of a low-frequency cut-off, so that peak intersymbol interference $\hat{U} = -1$. If rms intersymbol interference is held at one-quarter the peak value, i.e., $\underline{U} = 0.25$, the probability of encountering the maximum tolerable intersymbol interference and resultant errors in reception is low enough to be disregarded. On this basis the ratio ω_0/ω_1 would in accordance with (9.06) have to be less than 0.0625. Actually a substantially smaller ratio would be required because of intersymbol interference from other imperfections in the transmission characteristic and noise. Furthermore, a low-frequency cut-off will be accompanied by phase distortion at the low end of the transmission band, disregarded in the above evaluation. The requirements imposed on the low-frequency cut-off will thus be rather severe for a pulse system as assumed above in which random sequences of pulses are transmitted at intervals $\tau_1 = 1/2f_1$. Two pulse amplitudes were assumed above, and with a greater number of amplitudes the requirements would be more severe.

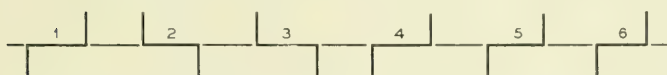
From Fig. 33 it is evident that the effect of a low-frequency cut-off on a received pulse train can be reduced by transmitting pulses at longer intervals than $\tau_1 = 1/2f_1$ considered above. For example, with a two-fold

increase in the pulse interval, as represented by the second case in Fig. 33, the maximum displacement of the zero line would be half the peak amplitude of the pulses. There would then be a 50 per cent reduction in the margin for distinction between the presence and absence of a pulse in a random pulse train, rather than a complete elimination of the margin for an infinite train of pulses of the same polarity transmitted at intervals $\tau_1 = 1/2f_1$. This improvement would be achieved at the expense of a two-fold increase in bandwidth for a given pulse transmission rate. A further improvement, for the same two-fold increase in bandwidth, can be achieved by "dipulse" transmission, as discussed below.

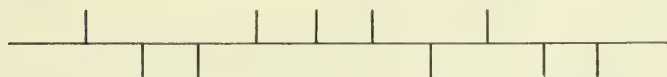
In dipulse transmission a positive pulse followed by a negative pulse in the next pulse position would be transmitted to indicate "on," and a negative pulse followed by a positive pulse to indicate "off," as indicated in Fig. 34. There will then be a substantial reduction in the pulse



(a) PULSE TRAIN WITH POSITIVE AND NEGATIVE IMPULSES



(b) PULSE TRAIN WITH POSITIVE AND NEGATIVE DIPULSES



(c) PULSE TRAIN (a) REVERSED AND DELAYED BY ONE PULSE INTERVAL



(d) DICODE TRANSMISSION (a)+(c)

THE PULSES AND ZEROS IN THE RECEIVED PULSE TRAIN (d) HAVE THE FOLLOWING RELATIONS TO THE ORIGINAL PULSES (a)

1. POSITIVE AND NEGATIVE PULSES IN (d) REPRESENT CORRESPONDING PULSES IN (a)
2. POINTS ON PULSE TRAIN IN (d) REPRESENT A REPETITION OF PREVIOUS PULSE, AS INDICATED BY DASHED LINES

Fig. 34 — Dipulse and dicode pulse transmission methods.

overlaps resulting from a low-frequency cut-off, as illustrated in Fig. 35, and in peak intersymbol interference.

If $\Delta P(t) = P(t) - P_0(t)$ is the modification in the impulse characteristic shown in Fig. 32, the modification in the dipulse transmission characteristic resulting from a low-frequency cut-off becomes

$$\Delta_1 P(t) = \Delta P(t) - \Delta P(t - \tau_1), \quad (9.07)$$

where τ_1 is the interval between the positive and negative dipulse components.

The difference given by (9.07) represents the differential in the curve $P(t) - P_0(t)$ shown in Fig. 32 over an interval τ_1 . It can be shown that the maximum cumulative effect or peak intersymbol interference for a long pulse train is represented by the sum of the differentials given by (9.07) and is approximately equal to

$$\hat{U} \cong \Delta P(\tau_1) = P(\tau_1) - P_0(\tau_1). \quad (9.08)$$

As an example, if the shape of $A - A_0$ in Fig. 32 were about the same as that of A_0 , $\Delta P(t)$ would have the same shape as $P_0(t)$ but would be lower in peak amplitude by the factor f_0/f_1 and would have the time scale increased by the factor f_1/f_0 . Peak intersymbol interference as

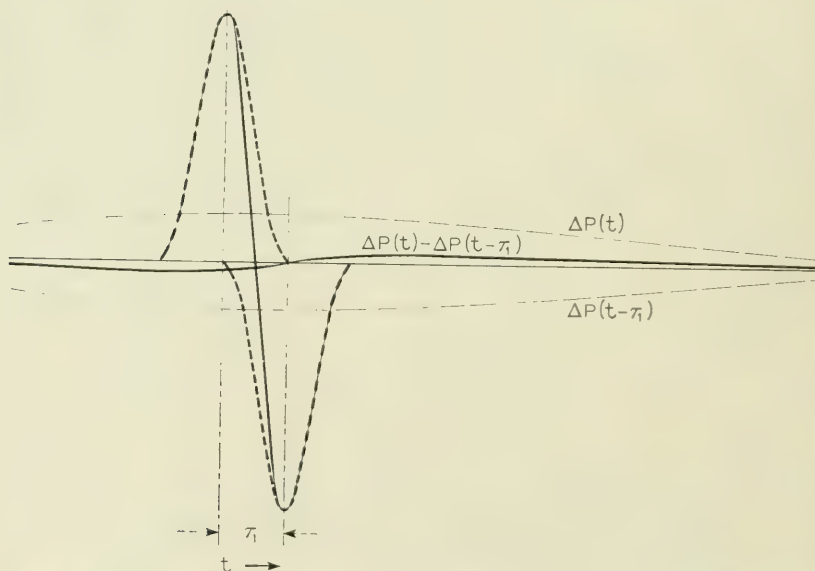


Fig. 35 — Low-frequency cut-off effects $\Delta P(t)$ and $\Delta P(t - \tau_1)$ for positive and negative pulses and resultant effect $\Delta_1 P(t) = \Delta P(t) - \Delta P(t - \tau_1)$ for a dipulse.

obtained from (9.08) would then be about $\hat{t} = f_0/f_1$ and thus in the order of 10 per cent of the peak pulse amplitude for $f_0/f_1 = 0.10$.

The bandwidth penalty incurred in dipulse transmission can be avoided by transmitting two identical pulse trains, one of which is delayed by one pulse interval and reversed in polarity with respect to the other.* The combined pulse train will then be as indicated in Fig. 34, and one or the other of the two original component pulse trains can be restored at the receiving end by suitable conversion equipment. In the combined pulse train, a pulse of one polarity is always followed by a pulse of opposite polarity, but not necessarily in the next pulse position. For this reason the low-frequency cut-off compensation with the above method of "dicode" transmission is not quite as effective as with dipulse transmission. Furthermore, since it is necessary to distinguish between three pulse amplitudes (1, 0, -1), in the received pulse train, the maximum tolerable pulse distortion in relation to the peak pulse amplitude is only half as great as with two pulse amplitudes (1, -1) in an ordinary code.

10. TRANSMISSION DISTORTION FROM BAND-EDGE PHASE DEVIATIONS

In pulse transmission systems where phase equalization is employed, it may be impracticable or unnecessary to equalize over the entire transmission band. There will then be residual phase distortion near the band-edges, as indicated in Fig. 36. This type of phase deviation will give rise to pulse distortion extending over appreciable time intervals if the band-edge phase deviations are large, as indicated in the above figure, for the reason that the frequency components outside the linear phase range will be received with increased transmission delay. Evaluation of the pulse shape is in this case a rather elaborate procedure, but rms pulse distortion or intersymbol interference resulting from such phase distortion can readily be determined as outlined below. In certain pulse modulation systems, such as PAM time division systems, rms intersymbol interference is of principal interest. In other systems where peak intersymbol interference is controlling, it may usually be estimated with engineering accuracy by applying a peak factor.

When the pulse shape is known, peak intersymbol interference may be determined by methods outlined in Section 13. Comparison of peak intersymbol interference evaluated in this manner with rms pulse distortion, for some cases in which the pulse shapes in the presence of phase

* L. A. Meacham originally proposed this method in an unpublished memorandum.

distortion were determined, indicates that the peak factor is about 3 when phase distortion is appreciable and the pulses are substantially prolonged in duration.

Returning to equation (8.17) and assuming $\alpha = 0$, the following relationship is obtained for rms intersymbol interference due to phase deviations

$$\underline{U} = \frac{1}{P(0)} \left(\frac{1}{\pi\tau_1} \right)^{1/2} \left[\int_0^\infty 2A_0^2 (1 - \cos \beta) d\omega \right]^{1/2}, \quad (10.01)$$

where $\beta = \beta(\omega)$ is the deviation from a linear phase characteristic.

For transmission systems with a linear phase shift, the peak amplitude of the pulses is given by (8.23) and with this relation in (10.01)

$$\underline{U} = \left(\frac{\pi}{\omega_1 \tau_1} \right)^{1/2} \lambda, \quad (10.02)$$

where

$$\lambda = \left[\omega_1 \int_0^{\omega_{\max}} 2A_0^2 (1 - \cos \beta) d\omega \right]^{1/2} / \left[\int_0^{\omega_{\max}} A_0 d\omega \right]. \quad (10.03)$$

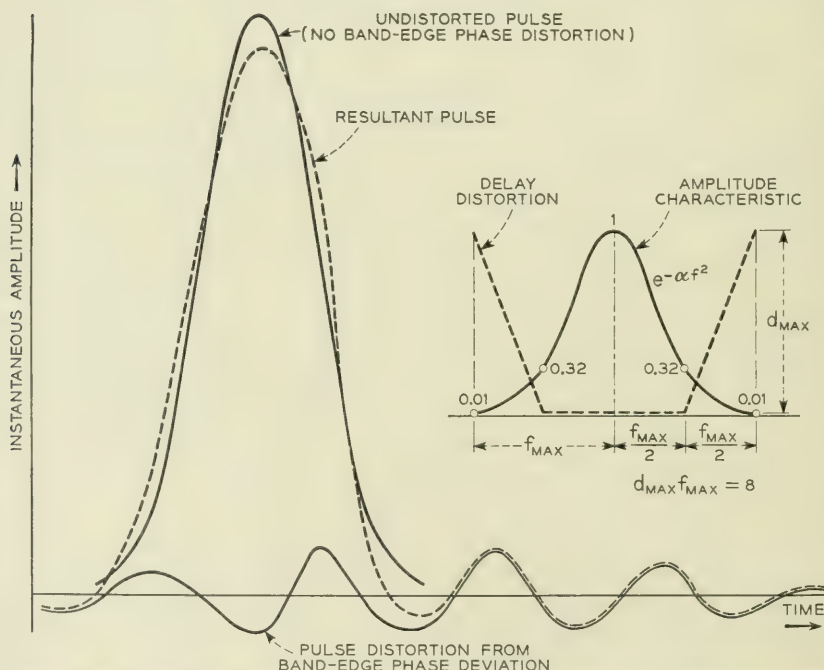


Fig. 36 — Pulse distortion from band-edge phase deviation for particular case of linear band-edge delay distortion.

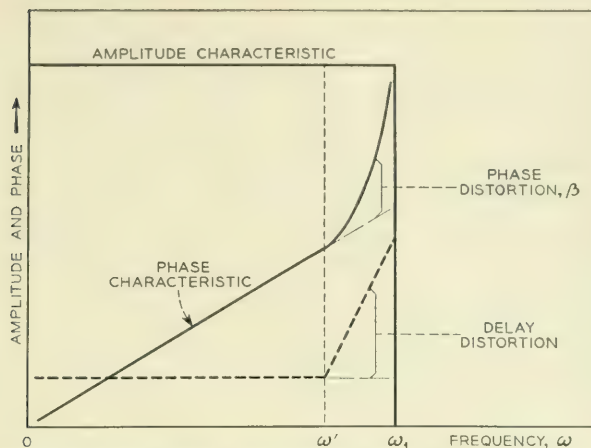


Fig. 37 — Constant amplitude characteristic with band-edge phase distortion.

If there is no phase distortion, i.e., $\beta = 0$, between $\omega = 0$ and ω' , equation (10.03) becomes

$$\lambda = \left[\omega_1 \int_{\omega'}^{\omega_{\max}} 2A_0^2 (1 - \cos \beta) d\omega \right]^{1/2} / \left[\int_0^{\omega_{\max}} A_0 d\omega \right]. \quad (10.04)$$

As an example, consider a parabolic deviation from a linear phase characteristic between ω' and $\omega_{\max}$, in which case delay distortion would vary linearly in this band, as indicated in Fig. 37 for a constant amplitude characteristic for which $\omega_{\max} = \omega_1$. In this case

$$\beta = \beta_1 \left(\frac{\omega - \omega'}{\omega_1 - \omega'} \right)^2, \quad (10.05)$$

where β_1 is the maximum phase deviation, obtained for $\omega = \omega_1$.

Equation (10.04) in the above case becomes:

$$\begin{aligned} \lambda^2 &= \frac{2}{\omega_1} \int_{\omega'}^{\omega_1} \left[1 - \cos \beta_1 \left(\frac{\omega - \omega'}{\omega_1 - \omega'} \right)^2 \right] d\omega, \\ &= \frac{2(\omega_1 - \omega')}{\omega_1} \left[1 - \beta_1^{-1/2} \int_0^{\beta_1^{1/2}} \cos u^2 du \right], \\ &= \frac{2(\omega_1 - \omega')}{\omega_1} \left[1 - \frac{1}{2} \left(\frac{\pi}{2\beta_1} \right)^{1/2} (R + X) \right], \end{aligned} \quad (10.06)$$

where $R + iX = \text{erf}(\beta_1^{1/2} e^{i\pi/4})$ in which erf is the error function.

For a constant amplitude transmission characteristic as assumed above, $(\pi/\omega_1\tau_1) = 1$, so that (10.02) becomes $U = \lambda$, which may also

be written:

$$\underline{U} = \left(\frac{\omega_1 - \omega'}{\omega_1} \right)^{1/2} \cdot F(\beta_1), \quad \text{and} \quad (10.07)$$

$$F(\beta_1) = 2^{1/2} \left[1 - \frac{1}{2} \left(\frac{\pi}{2\beta_1} \right)^{1/2} (R + X) \right]^{1/2}. \quad (10.08)$$

For various values of the maximum phase deviation β_1 in radians the function F becomes:

β_1	0	0.25	1	4	∞
F	0	0.14	0.43	1.24	1.42

If, for example, phase distortion were confined to 10 per cent of the transmission band, then $(\omega_1 - \omega')/\omega_1 = 0.1$. For a maximum phase deviation of 1 radian at the edge of the transmission band, $F = 0.43$ and $\underline{U} = 0.135$. For a maximum phase distortion of 4 radians, $F = 1.24$ and $\underline{U} = 0.39$. Since peak intersymbol interference may exceed the above rms values by a factor of about 3, and the maximum tolerable peak intersymbol interference in a system employing two pulse amplitudes would be less than 1, it is evident that band-edge phase deviations must be held at rather small values, less than about 3 radians, in the upper 10 per cent of the transmission band.

The above severe tolerances on band-edge phase distortion can be overcome by employing a transmission frequency characteristic of the type shown in Fig. 38 and previously discussed in Section 5. If the phase characteristic is linear between $\omega = 0$ and ω_1 , and phase distortion between ω_1 and $2\omega_1$ varies as

$$\beta = \beta_1 \left(\frac{\omega - \omega_1}{2\omega_1 - \omega_1} \right)^2 = \beta_1 \left(1 - \frac{\omega}{\omega_1} \right)^2, \quad (10.09)$$

equation (10.04) can be written

$$\begin{aligned} \lambda^2 &= \frac{1}{\omega_1} \int_{\omega_1}^{2\omega_1} \left(1 + \cos \frac{\pi\omega}{2\omega_1} \right)^2 \left[1 - \cos \beta_1 \left(1 - \frac{\omega}{\omega_1} \right)^2 \right] d\omega, \\ &= \int_0^1 \left(1 - \sin \frac{\pi}{2} u \right)^2 (1 - \cos \beta_1 u^2) du. \end{aligned} \quad (10.10)$$

Pulses may also in this case be transmitted at intervals $\tau_1 = \pi/\omega_1$ without intersymbol interference in the absence of phase distortion, so

that (10.02) becomes $\underline{U} = \lambda$ or

$$U = \left[\frac{1}{2} \int_0^1 \left(1 - \sin \frac{\pi}{2} u \right)^2 (1 - \cos \beta_1 u^2) du \right]^{1/2}. \quad (10.11)$$

The maximum delay distortion at the edge of the transmission band, i.e., $\omega = 2\omega_1$, is $d_{\max} = 2\beta_1/\omega_1$. The product of this delay distortion with the maximum frequency $f_{\max} = 2f_1$ is $d_{\max}f_{\max} = 2\beta_1/\pi$. For various values of maximum frequency β_1 and the corresponding product $d_{\max}f_{\max}$, the following phase values of rms intersymbol interference are obtained by numerical integration of (10.11). (This integral can be expressed in terms of a number of Fresnel integrals, but numerical integration is simpler and sufficiently accurate for the present purpose.)

β_1	π	2π	4π	∞
$d_{\max}f_{\max}$	2	4	8	∞
$\underline{U}$	0.070	0.120	0.185	0.330

The particular case $d_{\max}f_{\max} = 8$ is similar to that shown in Fig. 36, except that this figure applies to a Gaussian characteristic, for which the amplitude at $\omega = \omega_1$ has been taken as 0.32 rather than 0.5 in the case considered here. For this reason rms intersymbol interference from phase distortion would be greater in the present case.

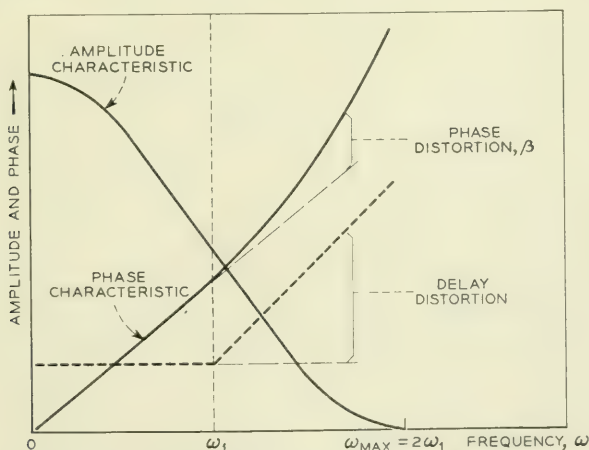


Fig. 38 — Typical transmission frequency characteristic with phase equalization over 50 per cent of transmission band.

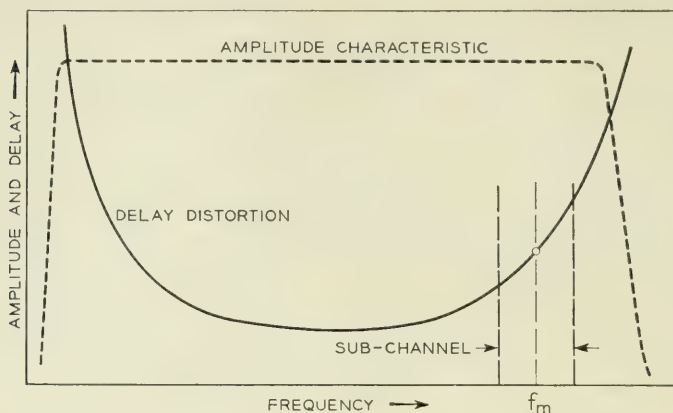


Fig. 39 — Sub-channel with nearly linear delay distortion.

Peak intersymbol interference may exceed the above rms values by a factor of about 3. In a system employing two pulse amplitude (1 and -1), the maximum tolerable intersymbol interference is 1. This value would thus be attained in the above case for $d_{\max} f_{\max} = \infty$. Hence, in a system employing two pulse amplitudes, and in the absence of noise and intersymbol interference from other sources, there would be no limitation on phase distortion for $\omega > \omega_1$, provided the phase characteristic is linear between $\omega = 0$ and ω_1 .

11. BAND-PASS CHARACTERISTICS WITH LINEAR DELAY DISTORTION

In Fig. 39 is shown a transmission frequency characteristic together with an assumed delay distortion $d\psi/d\omega$. When a portion of the transmission band is employed for pulse transmission, as for example in pulse signalling, data or telegraph transmission over portion of a voice channel, there may be an appreciable component of substantially linear delay distortion, as indicated in the above figure. The departure from a linear variation can usually be approximated by a cosine variation in delay, and the system can then be regarded as made up of two components in tandem, one with linear the other with cosine variation in delay. The effect of the latter can be evaluated by the methods outlined in Section 6, and the effect of a linear variation by methods established in this section.

In Fig. 40 is shown a symmetrical amplitude characteristic with linear delay distortion over the transmission band. Phase distortion with respect to the midband frequency is in this case

$$\Psi(u) = \beta u^2 \quad \text{and} \quad \Psi(-u) = \beta u^2, \quad (11.01)$$

and delay distortion

$$d\Psi(u)/du = 2\beta u, \quad d\Psi(-u)/du = -2\beta u. \quad (11.02)$$

(The symbol β , together with α , η , a and b used later in this section do not have the same meaning as in earlier sections.) With (11.01) in (2.10) and (2.11), the in-phase and quadrature components in (2.09) become

$$\begin{aligned} R_- + R_+ &= \frac{2\delta}{\pi} \int_0^\infty \alpha(u) \cos ut \cos \beta u^2, \quad \text{and} \\ Q_- + Q_+ &= \frac{2\delta}{\pi} \int_0^\infty \alpha(u) \cos ut \sin \beta u^2. \end{aligned} \quad (11.03)$$

The in-phase and quadrature components can accordingly be identified with the real and the negative imaginary component of the integral

$$J = \frac{2\delta}{\pi} \int_0^\infty \alpha(u) \cos ut e^{-i\beta u^2} du. \quad (11.04)$$

The solution of this integral is rather simple for the particular case of a Gaussian transmission characteristic

$$\alpha(u) = e^{-\alpha u^2}, \quad (11.05)$$

in which case

$$\begin{aligned} J &= \frac{2\delta}{\pi} \int_0^\infty e^{-(\alpha + i\beta)u^2} \cos ut du, \\ &= \frac{\delta}{[\pi(\alpha + i\beta)]^{1/2}} e^{-t^2/4(\alpha + i\beta)}. \end{aligned} \quad (11.06)$$

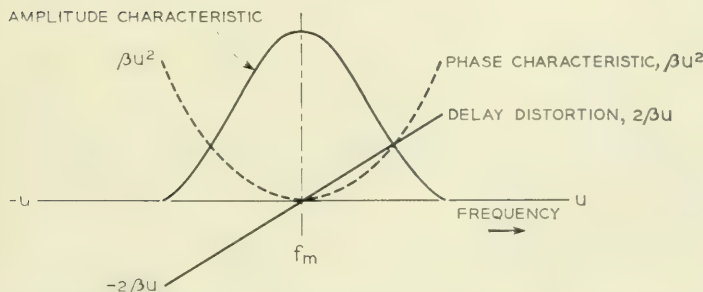


Fig. 40 — Symmetrical band-pass amplitude characteristic with linear delay distortion.

The real and negative imaginary components of this expression are

$$\begin{aligned} R_- + R_+ &= 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2} \cos(\Theta - bt^2), \quad \text{and} \\ Q_- - Q_+ &= 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2} \sin(\Theta - bt^2), \end{aligned} \quad (11.07)$$

where

$$\begin{aligned} a &= \frac{\alpha}{4(\alpha^2 + \beta^2)} & b &= \frac{\beta}{4(\alpha^2 + \beta^2)} & c &= (\alpha^2 + \beta^2)^{1/2} \\ \tan 2\Theta &= \beta/\alpha = b/a \end{aligned}$$

The impulse characteristic obtained with (11.07) in (2.09) becomes

$$\begin{aligned} P(t) &= 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2} [\cos(\omega_r t - \psi_r) \cos(\Theta - bt^2) \\ &\quad + \sin(\omega_r t - \psi_r) \sin(\Theta - bt^2)]. \end{aligned} \quad (11.08)$$

From (11.08) it is seen that the envelope is

$$\bar{P}(t) = 2\delta \left(\frac{c}{\pi}\right)^{1/2} e^{-at^2}. \quad (11.09)$$

The peak of the envelope obtained with $t = 0$ is smaller than without delay distortion ($\beta = 0$) by the factor

$$\eta = \frac{1}{[1 + (\beta/\alpha)^2]^{1/2}}. \quad (11.10)$$

The constant a is smaller than without delay distortion by the factor η^4 . If t_0 designates the time required for the instantaneous amplitude of a pulse to decay from its peak to a given value without delay distortion, the time t_1 to reach the same amplitude with delay distortion is

$$t_1 = t_0/\eta^2 = t_0[1 + (\beta/\alpha)^2]^{1/2}. \quad (11.11)$$

If $\omega_{\max}$ indicates the frequency at the 40 db down point on the transmission frequency characteristic, $\alpha\omega_{\max}^2 = 4.6$. The corresponding delay distortion is $d_{\max} = 2\omega_{\max}\beta$. Thus $\beta/\alpha = .68 d_{\max}f_{\max}$ so that (11.11) becomes:

$$t_1 = t_0[1 + 0.46 (d_{\max}f_{\max})^2]^{1/2}. \quad (11.12)$$

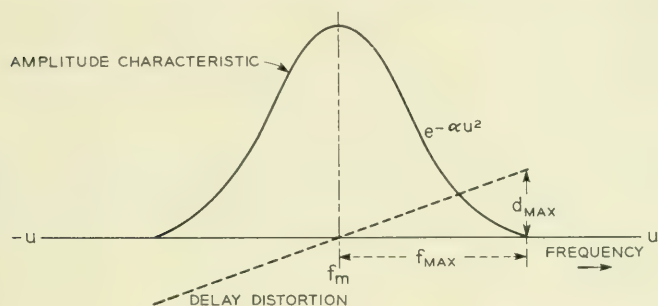
The effect of a linear delay distortion across the transmission band is thus to disperse or broaden the envelope of the received pulses, as illus-

trated in Fig. 41. For a specified pulse overlap or intersymbol interference the pulse spacing must accordingly be increased by the factor t_1/t_0 , so that for a given transmission performance the transmission capacity is reduced by the factor t_0/t_1 . About the same effect would be expected for other pulse shapes or amplitude characteristics resembling the Gaussian shape assumed in the above derivation.

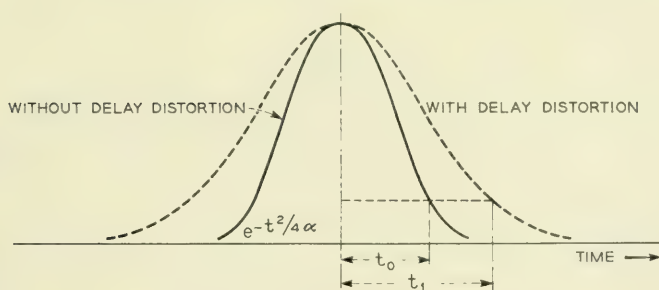
Comparison of (11.08) with (2.13) shows that the function $\varphi(t)$ with respect to the midband frequency is

$$\varphi(t) = \theta - bt^2. \quad (11.13)$$

If the reference or carrier frequency is displaced from the midband by



(a) TRANSMISSION FREQUENCY CHARACTERISTIC



(b) IMPULSE CHARACTERISTICS WITHOUT AND WITH DELAY DISTORTION

$$t_1 = t_0 [1 + 0.46(d_{MAX} f_{MAX})^2]^{\frac{1}{2}}$$

f_{MAX} = FREQUENCY FROM MIDBAND AT WHICH AMPLITUDE OF TRANSMISSION FREQUENCY CHARACTERISTIC IS REDUCED 40 DECIBELS

d_{MAX} = DELAY DISTORTION AT f_{MAX} IN SECONDS

Fig. 41 — Lengthening of impulse envelope by linear delay distortion for Gaussian transmission characteristic.

ω_y , the in-phase and quadrature components are in accordance with (2.18)

$$\begin{aligned} R_-' + R_+' &= \cos(\theta - bt^2 + \omega_y t - \psi_y) \bar{P}(t), & \text{and} \\ Q_-' - Q_+' &= \sin(\theta - bt^2 + \omega_y t - \psi_y) \bar{P}(t), \end{aligned} \quad (11.14)$$

where $\psi_y = \beta\omega_y^2$ and $\bar{P}(t)$ is given by (11.09).

REFERENCES

1. Y. W. Lee, Synthesis of Electric Networks by Means of the Fourier Transforms of Laguerre's Functions, *J. Math. & Phys.*, June, 1932.
2. H. W. Bode, *Network Analysis and Negative Feedback Amplifier Design*, D. Van Nostrand Book Company, 1945.
3. H. Nyquist, Certain Topics in Telegraph Transmission Theory, *A.I.E.E. Trans.*, April, 1928.
4. H. Nyquist and K. W. Pfeiffer, Effect of Quadrature Component in Single Sideband Transmission, *B.S.T.J.*, Jan., 1940.
5. H. A. Wheeler, The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *Proc. I.R.E.*, June, 1939.
6. C. R. Burrows, Discussion of 5 above, *Proc. I.R.E.*, June, 1939.
7. G. N. Watson, *Theory of Bessel Functions*, Cambridge University Press, 1944.
8. E. Jahnke and F. Emde, *Funktionentafeln*, 1928, p. 149.
9. E. A. Guillemin, *Communication Networks*, John Wiley & Sons, Inc., 1935.
10. W. R. Bennett, Time Division Multiplex Systems, *B.S.T.J.*, April, 1941.
11. S. Goldman, *Frequency Analysis, Modulation and Noise*, McGraw-Hill Book Co., 1948.
12. C. E. Shannon, A Mathematical Theory of Communication, *B.S.T.J.*, October, 1948.
13. B. M. Oliver, J. R. Pierce and C. E. Shannon, The Philosophy of PCM, *Proc. I.R.E.*, Nov., 1948.

Bell System Technical Papers Not Published in this Journal

AIKENS, A. J.,¹ and C. S. THAELEL.²

Noise and Crosstalk Control on N1 Carrier Systems, Elec. Eng., **72**, pp. 1075-1080, Dec., 1953.

ALLEY, R. E., JR.,¹ and F. J. SCHNETTLER.¹

Effect of Cross-Section Area and Compression Upon the Relaxation in Permeability for Toroidal Samples of Ferrites, Letter to the Editor, J. Appl. Phys., **24**, pp. 1524-1525, Dec., 1953.

ALLIS, W. P.,¹ and D. J. ROSE.¹

The Transition From Free to Ambipolar Diffusion, Phys. Rev., **93**, p. 84-93, Jan. 1, 1954.

ANDERSON, O. L.,¹ and D. A. STUART.⁴

Statistical Theories as Applied to the Glassy State, Ind. Eng. Chem., **46**, pp. 154-160, Jan., 1954.

ARNOLD, S. M., see S. E. KOONCE.

BARSTOW, J. M.,¹ and H. N. CHRISTOPHER.¹

The Measurement of Random Monochrome Video Interference. A.I.E.E., Commun. and Electronics, pp. 735-741, Jan., 1954.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

⁴ Cornell University.

BASHKOW, T. R.¹

Stability Analysis of a Basic Transistor Switching Circuit, Proc. National Electronics Conference, **9**, p. 748, Feb. 15, 1954.

BLECHER, F. H.¹

Automatic Gain Control of Junction Transistor Amplifiers, Proc. National Electronics Conference, **9**, p. 731, Feb. 15, 1954.

BOGERT, B. P.¹

Erratum: On the Band Width of Vowel-Formants, [published in J. Acous. Soc. Am., **25**, p. 791, (1953)], J. Acous. Soc. Am., **25**, p. 1203, Nov., 1953. The statement in the abstract which reads "The mean values for bars 1, 2, and 3 were 130, 100, and 185 cps, respectively," should be corrected to read "The median values for bars 1, 2, and 3, were 130, 150, and 185 cps, respectively."

BROWN, W. L.,¹ R. C. FLETCHER,¹ and K. A. WRIGHT.⁵

Annealing of Bombardment Damage in Germanium — Experimental, Phys. Rev., **92**, pp. 591–596, Nov. 1, 1953.

BROWN, W. L., see R. C. FLETCHER.

BURTON, J. A.,¹ G. W. HULL,¹ F. J. MORIN,¹ and J. C. SEVERIENS.¹

Effect of Nickel and Copper Impurities on the Recombination of Holes and Electrons in Germanium, J. Phys. Chem., **57**, pp. 853–859, Nov., 1953.

BURTON, J. A.,¹ R. C. PRIM,¹ and W. P. SLICHTER.¹

Distribution of Solute in Crystals Grown from the Melt — Theoretical, J. Chem. Phys., **21**, pp. 1987–1991, Nov., 1953.

BURTON, J. A.,¹ E. D. KOLB,¹ W. P. SLICHTER,¹ and J. D. STRUTHERS.¹

Distribution of Solute in Crystals Grown from the Melt — Experimental, J. Chem. Phys., **21**, pp. 1991–1996, Nov., 1953.

CAMPBELL, M. E., see C. L. LUKE.

¹ Bell Telephone Laboratories, Inc.

⁵ Massachusetts Institute of Technology.

CASE, R. L.,¹ and IDEN KERNEY.¹

Program Transmission Over Type-N Carrier Telephone, A.I.E.E., Commun. and Electronics, pp. 791-795, Jan., 1954.

CHRISTOPHER, H. N., see J. M. BARSTOW.

CLARK, M. A.¹

An Acoustic Lens as a Directional Microphone, J. Acous. Soc. Am., **25**, pp. 1152-1153, Nov., 1953.

CORENZWIT, E., see S. GELLER.

COY, J. A.¹

Heat Dissipation from Toll Transmission Equipment, A.I.E.E., Commun. and Electronics, pp. 762-768, Jan., 1954.

DUNN, H. K.¹

Remarks on a Paper Entitled "Multiple Helmholtz Resonators," Letter to the Editor, J. Acous. Soc. Am., **26**, p. 103, Jan., 1954.

FLETCHER, R. C.,¹ and W. L. BROWN.¹

Annealing of Bombardment Damage in a Diamond-Type Lattice — Theoretical, Phys. Rev., **92**, pp. 585-590, Nov. 1, 1954.

FLETCHER, R. C., see W. L. BROWN.

FRACASSI, R. D.,¹ and H. KAHL.¹

Type ON Carrier Telephone, A.I.E.E., Commun. and Electronics, pp. 713-721, Jan., 1954.

FULLER, C. S., see J. R. SEVERIENS.

GELLER, S.,¹ and E. CORENZWIT.¹

Hafnium Oxide, HfO₂ (Monoclinic), Anal. Chem., **25**, p. 1774, Nov., 1953.

¹ Bell Telephone Laboratories, Inc.

GIBBS, W. B.³

Investment Cast Artificial Larynx Eases Fabrication Difficulties, Precision Metal Molding, pp. 34, 35, 75, Dec., 1953.

GREEN, F. O.³

Cost Control — Bonus from Centralized Maintenance, Plant Engineering, pp. 88-91, Jan., 1954.

HAGSTRUM, H. D.¹

Instrumentation and Experimental Procedure for Studies of Electron Ejection by Ions and Ionization by Electron Impact, Rev. Sci. Instr., **24**, pp. 1122-1142, Dec., 1953.

HANSON, A. N.¹

Automatic Testing of Wired Relay Circuits, A.I.E.E., Commun. and Electronics, pp. 805-857, Jan., 1954.

HULL, G. W., see J. A. BURTON.

KAHL, H., see R. D. FRACASSI.

KERNEY, IDEN, see R. L. CASE.

KOCH, W. E.¹

Use of the Sound Spectrograph for Appraising the Relative Quality of Musical Instruments, Letter to the Editor, J. Acous. Soc. Am., **26**, p. 105, Jan., 1954.

KOLB, E. G., see J. A. BURTON.

KOONCE, S. E.¹ and S. M. ARNOLD.¹

Metal Whiskers, Letter to the Editor, J. Appl. Phys., **25**, pp. 134-135, Jan., 1954.

¹ Bell Telephone Laboratories, Inc.

³ Western Electric Company, Inc.

KRETZMER, E. R.¹

An Amplitude-Stabilized Transistor Oscillator, Proc. National Electronics Conference, p. 756, Feb. 15, 1954.

LEWIS, H. W.¹

Search for the Hall Effect in a Super-Conductor — Experiment, Phys. Rev., **92**, pp. 1149–1151, Dec., 1953.

LINVILLE, J. G.¹

A New RC Filter Employing Active Elements, Proc. National Electronics Conference, **9**, p. 342, Feb. 15, 1954.

LUKE, C. L.,¹ and M. E. CAMPBELL.¹

Determination of Impurities in Germanium and Silicon, Anal. Chem., **25**, pp. 1588–1593, Nov., 1953.

MAHONEY, J. J., see E. H. PERKINS.

MAY, J. E.¹

Characteristics of Ultrasonic Delay Lines Using Quartz and Barium Titanite Ceramic Transducer, Proc. National Electronics Conference, **9**, p. 264, Feb. 15, 1954.

MORIN, F. J.¹

Lattice Scattering Mobility in Germanium, Phys. Rev., **93**, pp. 62–63, Jan. 1, 1954.

MORIN, F. J., see J. A. BURTON.

MURPHY, E. J.¹

Surface Migration of Water Molecules in Ice, J. Chem. Phys., **21**, pp. 1831–1835, Oct., 1953.

PENNELL, E. S.¹

A Temperature Controlled Ultrasonic Solid Delay Line, Proc. National Electronics Conference, **9**, p. 255, Feb. 15, 1954.

¹ Bell Telephone Laboratories, Inc.

PERKINS, E. H.,¹ and J. J. MAHONEY.¹

Type-N Carrier Telephone Deviation Regulator, A.I.E.E., Commun. and Electronics, pp. 757-762, Jan., 1954.

PETERSON, G. E.,⁶ and GORDON RAISBECK.¹

The Measurement of Noise with the Sound Spectrograph, J. Acous. Soc. Am., **25**, p. 1157, Nov., 1953.

PRIM, R. C., see J. A. BURTON.

PRINCE, M. B.¹

Drift Mobilities in Semi-Conductors. I — Germanium, Phys. Rev., **92**, pp. 681-687, Nov. 1, 1953.

RAISBECK, GORDON, see G. E. PETERSON.

REA, W. W.³

Oscilloscope Reduces Cost of Jig Grinding Operations, Machinery, pp. 201-203, Dec., 1953.

REMEIKA, J. P.

Method for Growing Barium Titanate Single Crystals, J. Am. Chem. Soc., **76**, pp. 940-941, Feb. 5, 1954.

ROSE, D. J., see W. P. ALLIS.

SCHLAACK, N. F.¹

Development of the LD Radio System, Trans. I.R.E., Professional Group on Communication Systems, pp. 29-38, Jan., 1954.

SCHNETTLER, F. J., see R. E. ALLEY, JR.

SEVERIENS, J. C., see J. A. BURTON.

¹ Bell Telephone Laboratories, Inc.

³ Western Electric Company, Inc.

⁶ University of Michigan.

SEVERIENS, J. R.,¹ and C. S. FULLER.¹

Mobility of Impurity Ions in Germanium and Silicon, Letter to the Editor, *Phys. Rev.*, **92**, pp. 1322, Dec., 1953.

SHOCKLEY, W.¹

Transistor Physics, *Am. Scientist*, **42**, pp. 41-72, Jan., 1954.

SHOCKLEY, W.¹

Some Predicted Effects of Temperature Gradient on Diffusion in Crystals, Letter to the Editor, *Phys. Rev.*, **93**, pp. 345-346, Jan. 15, 1954.

SLICHTER, E. D., see J. A. BURTON.

SNOREK, F.³

Cut Tool Costs with Precision Casting, *Iron Age*, pp. 136-139, Feb. 11, 1954.

STILES, K. P.²

Overseas Radio Telephone Services of A. T. and T. Co., *Trans. I.R.E., Professional Group Communications Systems*, pp. 39-44, Jan., 1954.

STRUTHERS, J. D., see J. A. BURTON, and C. D. THURMOND.

STUART, D. A., see O. L. ANDERSON.

THAELER, C. S., see A. J. AIKENS.

THURMOND, C. D.¹

Equilibrium Thermochemistry of Solid and Liquid Alloys of Germanium and Silicon — The Solubility of Ge and Si in Elements of Group III, IV and V, *J. Phys. Chem.*, **57**, pp. 827-830, Nov., 1953.

THURMOND, C. D.,¹ and J. D. STRUTHERS.¹

Equilibrium Thermochemistry of Solid and Liquid Alloys of Ger-

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

³ Western Electric Company, Inc.

manium and of Silicon — The Retrograde Solid Solubilities of Sb in Ge, Cu in Ge, and Cu in Si, *J. Phys. Chem.*, **57**, pp. 831-834, Nov., 1953.

WALKER, L. R.¹

Dispersion Formula for Plasma Waves, Letter to the Editor, *J. Appl. Phys.*, **25**, pp. 131-132, Jan., 1954.

WOLFF, P. A.¹

Theory of Plasma Waves in Metals, *Phys. Rev.*, **92**, pp. 18-23, Oct. 1, 1953.

WRIGHT, K. A., see W. L. BROWN.

¹ Bell Telephone Laboratories, Inc.

Contributors to this Issue

M. M. ATALLA, B.S., Cairo University, 1945; M.S., Purdue University, 1947; Ph.D., Purdue University, 1949; Studies at Purdue undertaken as the result of a scholarship from Cairo University for four years of graduate work. Bell Telephone Laboratories, 1950-. For the past three years he has been a member of the Switching Apparatus Development Department, in which he is supervising a group doing fundamental research work on contact physics and engineering. Current projects include fundamental studies of gas discharge phenomena between contacts, their mechanisms, and their physical effects on contact behavior; also fundamental studies of contact opens and resistance. In 1950, an article by him was awarded first prize in the junior member category of the A.S.M.E. He is a member of Sigma Xi, Sigma Pi Sigma, Pi Tau Sigma, the American Physical Society, and an associate member of the A.S.M.E.

JAMES M. EARLY, B.S., cum laude, New York State College of Forestry, 1943; M.S. and Ph.D. Ohio State University, 1948 and 1951. Bell Telephone Laboratories 1951-. After teaching Electrical Engineering at Ohio State University for five years while studying for his Master's and Ph.D. degrees, Dr. Early joined an electronic apparatus development group, participating in the development of the junction transistor. At present he is doing general theoretical work as well as development work on high frequency junction transistors. Member of the I.R.E. and Eta Kappa Nu. Associate of Sigma Xi.

WALTER T. EPPLER, B.S., in E.E., Tufts College, 1927; Western Electric Company, purchase of special machinery and development of coil manufacture, 1927-1932; Crystal Golf Ball Company, Superintendent of Manufacture, 1933-1936; New Haven Clock Company, 1936-1937; Western Electric Company, 1937-. In 1941, he engineered a conveyORIZED dispatch control system for key assembly at the Kearny plant. For the past six years, he has been engaged in cable sheath engineering on Alpth and Stalpth cable. Member of New Jersey Society of Professional Engineers.

STEWART E. MILLER, University of Wisconsin, 1936-39; B.S. and M.S., Massachusetts Institute of Technology, 1941. Bell Telephone Laboratories, 1941-. Except for World War II work on airborne radar systems, Mr. Miller's first eight years at the Laboratories were concerned with studies on coaxial carrier transmissions systems. A member of the radio research group, he is currently in charge of research on guided systems and associated millimeter and microwave techniques at Holmdel. Member of the I.R.E., Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

HARRY SUHL, B.Sc., University of Wales, 1943; Ph.D., Oriel College, University of Oxford, 1948. Admiralty Signal Establishment, 1943-46; Bell Telephone Laboratories, 1948-. Dr. Suhl conducted research on the properties of germanium until 1950 when he became concerned with electron dynamics and solid state physics research. His current work is in the applied physics of solids. Member of the American Institute of Physics and Fellow of the American Physical Society.

ERLING D. SUNDE, E.E., Technische Hochschule, Darmstadt, Germany, 1926. Brooklyn Edison Company, 1927; American Telephone and Telegraph Company, 1927-1934; Bell Telephone Laboratories, 1934-. Mr. Sunde's work has been centered on theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series Book. Member of the A.I.E.E., the American Mathematical Society, and the American Association for the Advancement of Science.

LAURENCE R. WALKER, B.Sc. and Ph.D., McGill University, 1935 and 1939; University of California, 1939-41. Radiation Laboratory, Massachusetts Institute of Technology, 1941-1945; Bell Telephone Laboratories, 1945-. Dr. Walker has been primarily engaged in research on microwave oscillators and amplifiers. At present he is a member of the physical research group concerned with the applied physics of solids. Fellow of the American Physical Society.

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII

JULY 1954

NUMBER 4

KANSAS CITY, MO.

Negative Resistance Arising from Transit Time in Semiconductor Diodes W. SHOCKLEY 799

Transistor and Junction Diodes in Telephone Power Plants F. H. CHASE, B. H. HAMILTON AND D. H. SMITH 827

Wire Straightening and Molding for Wire Spring Relays A. J. BRUNNER, H. E. COSSON AND R. W. STRICKLAND 859

Some Fundamental Problems in Percussive Welding E. E. SUMNER 885

Automatic Contact Welding in Wire Spring Relay Manufacture A. L. QUINLAN 937

Electronic Relay Tester T. E. DAVIS AND A. L. BLAHA 925

Topics in Guided Wave Propagation Through Gyromagnetic Media
Part II — Transverse Magnetization and Non-Reciprocal Helix
H. SUHL AND L. R. WALKER 939

Theoretical Fundamentals of Pulse Transmission — II
E. D. SUNDE 987

Bell System Technical Papers Not Published in this Journal 1011

Recent Bell System Monographs 1017

Contributors to this Issue 1019

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

S. BRACKEN, *Chairman of the Board,*
Western Electric Company

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. McNEELY, *Vice President, American Telephone
and Telegraph Company*

EDITORIAL COMMITTEE

W. H. DOHERTY, <i>Chairman</i>	F. R. LACK
A. J. BUSCH	W. H. NUNN
G. D. EDWARDS	H. I. ROMNES
J. B. FISK	H. V. SCHMIDT
E. I. GREEN	G. N. THAYER
R. K. HONAMAN	J. R. WILSON

EDITORIAL STAFF

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXIII

JULY 1954

NUMBER 4

Copyright, 1954, American Telephone and Telegraph Company

Negative Resistance Arising from Transit Time in Semiconductor Diodes

By W. SHOCKLEY

(Manuscript received January 22, 1954)

The structural simplicity of two-terminal compared to three-terminal devices indicates the potential importance of two terminal devices employing semiconductors and having negative resistance at frequencies properly related to the transit time of carriers through them. Such negative resistances may be combined with unsymmetrically transmitting components, such as gyrators or Hall effect plates, to form dissected amplifiers that may be made to simulate conventional three-terminal amplifiers and operate at high frequencies. The characteristics of several structures are analyzed on the basis of theory and it is found that negative resistances are possible for properly designed structures.

1. NEGATIVE RESISTANCE AND DISSECTED AMPLIFIERS

Because the drift velocities of current carriers in semiconductors are smaller than the velocities attainable in vacuum tubes, transistor structures must be smaller to achieve comparable frequencies. In principle it is possible, of course, to make compositional structures (i.e., distributions of donors and acceptors) in semiconductor crystals on a scale much smaller than is possible for vacuum tubes. At present, however, the available techniques are limited and it may require many years before the ultimate potentialities are approached.

It is instructive, however, to speculate on some of these ultimate potentialities. For example a grain boundary formed of edge type dislocations is in a sense an analogue of a grid. Possibly it can be made into a grid by acting as a locus for an atmosphere of donors or acceptors. Evidently such a grid will approach the smallest spacing that can be achieved with any known form of matter. If the spacings perpendicular to the grid are made comparable to a mean free path of the carriers used, the device will operate like a vacuum tube with carrier velocities controlled by inertia rather than by mobility. It is not easy to conceive of a structure having the potentiality of operating at higher frequencies.

It is evident that the difficulty of making small structures increases with the number of electrodes. For example, it is now possible to make diodes which give usable rectification at frequencies above 10^{10} cps. In these the "working volume" is a very thin layer under the metal point. The thickness of this layer is controlled by surface treatments and the applied voltages. The diameter of the point, which is the minimum dimension mechanically controlled, is much larger than this thickness of the layer. In order to make a transistor of comparable frequency, it would be necessary to make structural elements having dimensions comparable to the thickness of the layer and this would be a much more exacting task than making the diode.

These considerations point out the importance of giving serious consideration to two-terminal structures as amplifying elements. It is possible, in principle at least, to have structures which are much smaller in one dimension than the other two and which exhibit negative resistance and thus give ac power at frequencies comparable to the reciprocal of the transit time across the small dimension.

The attractiveness of such negative resistance diodes for amplification is enhanced by the possibility of using them in *dissected amplifiers*^{1,1} in combination with nonreciprocal elements such as gyrators or Hall effect plates. Combinations of negative resistance elements and nonreciprocal elements can lead to structures having gain and unsymmetrical transmission that simulate conventional amplifiers. The adjective *dissected* has been suggested for them since elements giving power gain are physically separated from those giving one-way transmission.

In this article we shall not consider the possible forms of dissected amplifiers, of which there are a wide variety. Instead we shall give an introductory treatment of some forms of negative resistance that may arise from transit time effects. In some cases the most instructive way of treating the structure is by way of the "impulsive impedance" and we devote most of the next section to considering this method.

2. THE IMPULSIVE IMPEDANCE AND NEGATIVE RESISTANCE

The impulsive impedance $D(t)$ for a two terminal device is defined in terms of its transient response to an impulse of current. Thus if the current through the device is

$$J(t) = J + j(t),$$

where J is the dc current and

$$j(t) = 0$$

except very near $t = 0$ and

$$\int j(t) dt = \delta Q,$$

then the voltage is

$$V(t) = V + v(t),$$

where V is the dc voltage and

$$v(t) = \delta Q D(t).$$

In other words, if in addition to the dc biasing current, a charge δQ is instantaneously forced through the circuit at time $t = 0$, the added voltage is $D(t)$ per unit charge. These equations also serve to introduce the notation used in this article:

Notation. In general, quantities that are functions of time or position will have the functional dependence explicitly indicated. In Sections 4 and 5, however, the symbol δ will be used to distinguish the transient parts δE and $\delta \rho$ from the dc parts of the electric field and charge density.

Capital $V(t)$ and $J(t)$ stand for total voltages and current. Without functional dependence upon (t) they are the dc parts. Similarly $v(t)$ and $j(t)$ are the ac or transient parts. A sinusoidal disturbance is represented by

$$v(t) = v \exp i\omega t,$$

$$j(t) = j \exp i\omega t,$$

where v and j are not functions of time. Where it is necessary to distinguish the displacement current at a particular location from the conduction current, as in the next section, we shall write

$$j(D, S_2, t),$$

meaning the displacement current across space charge region S_2 as a function of time.

In this section we shall treat J and j as circuit currents. In subsequent sections, we shall be concerned with current densities and shall use the same symbols.

The complex impedance of the device is evidently

$$Z(\omega) = v/j,$$

where v and j are the coefficients in the sinusoidal case.

In terms of the system of notation introduced above, $Z(\omega)$ may also be expressed in terms of $D(t)$ by expressing $j \exp i\omega t$ in terms of increments of charge

$$dQ = j e^{i\omega t} dt,$$

and summing over all increments up to time t . This leads to

$$Z(\omega) = \int_0^\infty D(t) \exp(-i\omega t) dt.$$

A negative resistance will occur if

$$0 > \int_0^\infty D(t) \cos \omega t dt = (-1/\omega) \int_0^\infty D'(t) \sin \omega t dt,$$

the latter form coming from integration by parts for the case of $D(\infty) = 0$, the only situation treated in this article.

The use of $D(t)$ in analysing the potential merits of diode structures from the point of view of negative resistance is illustrated in Fig. 2.1. Here three cases of $D(t)$ together with certain cosine waves are shown. It is seen for case (a) that a negative real part of Z will be obtained. For case (b), the real part of Z is zero for the frequency shown; this represents a limit; for other frequencies, a positive real part will be ob-

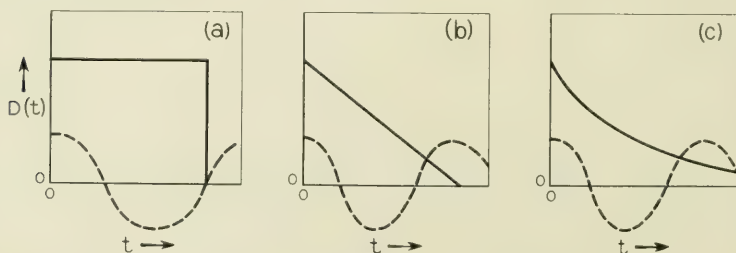


Fig. 2.1 — Some hypothetical $D(t)$ characteristics.

tained. Case (c) represents an exponential fall such as might occur for a capacitor and resistor in parallel. We shall discuss this example below.

The conclusions regarding (a) and (b) may be somewhat more easily seen from the corresponding $-D'(t)$ plots shown in Fig. 2.2. From part (a) it can be seen that the negative maximum in the sine wave at the end of the rectangular $D(t)$ plot is particularly favorable. From part (b) it is seen that no choice of ω will result in more negative area of sine wave than positive. For (c) it is evident that each positive half cycle of the sine wave gives a larger contribution than the following negative half cycle and hence that a positive resistance will be obtained.

For case (c), it is instructive to obtain the value of Z analytically by using

$$D(t) = C^{-1} \exp(-t/RC).$$

This leads correctly to

$$Z(\omega) = (R^{-1} + i\omega C)^{-1}.$$

For small values of ωRC , Z reduces to R ; furthermore, for this case, the decay of $D(t)$ occurs while $\cos \omega t = 1$. Under these conditions

$$Z(\omega) = \int_0^{\infty} D(t) dt.$$

This result is useful for estimating the effect of quickly decaying contributions to $D(t)$. These evidently contribute a positive resistance to Z equal to the area under the $D(t)$ curve.

From these considerations it follows that an upward deviation from the linear fall in Fig. 2.1(b) towards Fig. 2.1(a) will result in negative resistance. In Sections 4 and 5 we shall see how particular structures may lead to such favorable, convex-upwards characteristics for $D(t)$.

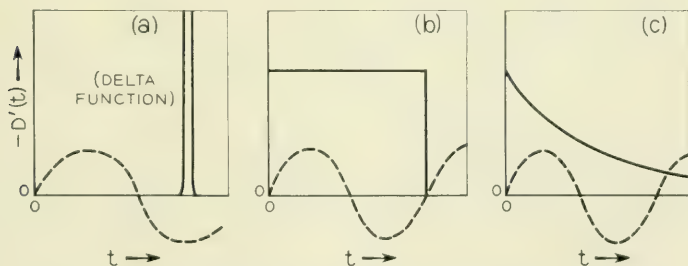


Fig. 2.2 — The $-D'(t)$ characteristic corresponding to fig. 2.1.

3. MINORITY CARRIER DELAY DIODE

As a first example we shall consider the behavior of the device shown in Fig. 3.1. We have chosen a p-n-p structure rather than an n-p-n so as to deal with positively charged carriers and thus avoid numerous minus signs in the equations. In this figure we have used capital letters P and N to designate specific regions, reserving the small letters to indicate carrier densities and conductivity types.

Several features that simplify the theoretical treatment should be noted:

- (a) The P_1N junction is 100-fold more heavily doped on the P_1 -side.
- (b) The doping in the layer N varies exponentially with distance by a factor of 10 across the layer.

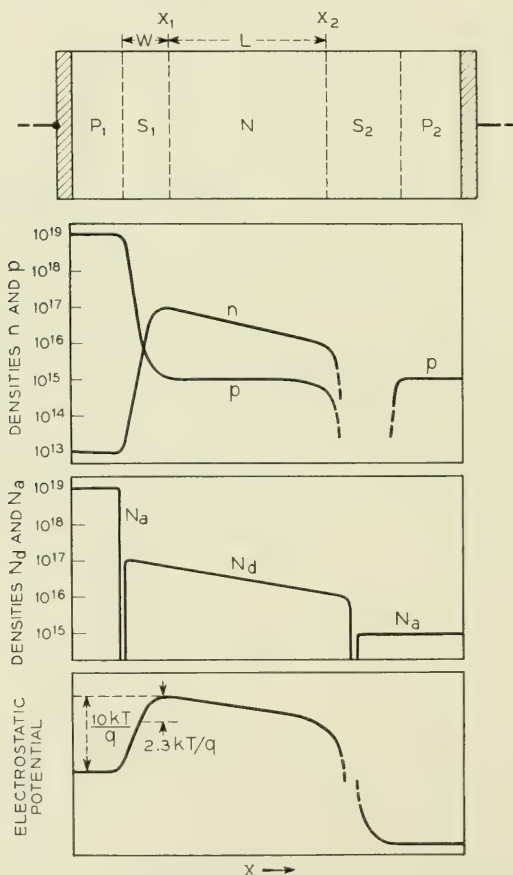


Fig. 3.1 — Constitution of minority carrier delay diode.

(c) Throughout N the concentration of holes is less by a factor of 10 than the electron concentration.

(d) The thickness of N is large compared to the depth of space charge penetration into it.

(e) The voltage drop across the space charge region S_2 is large compared to the other voltage drops.

The conditions lead at the operating frequency to the following consequences:

(1) The current across the first junction is carried preponderantly by holes.

(2) The hole drift in N is substantially unaffected by the ac field and thus represents the delayed diffusing and drifting current injected across the first junction.

(3) The ac voltage drop occurs chiefly across S_2 .

We shall show below (Section 3.2) how (a) to (e) lead to (1) to (3), but we shall first bring out the importance of (1) to (3) by using them to give a simple treatment of the impedance of the diode.

3.1. Calculation of Impedance.

If the total current is

$$J(t) = J + je^{i\omega t}, \quad (3.1)$$

then the ac hole current across S_1 is also in the notation discussed in Section 2 with the addition of the symbol p to indicate holes

$$j(p, S_1, t) = je^{i\omega t}. \quad (3.2)$$

This current flows through the n-layer unaffected by the ac field and arrives at S_2 delayed and attenuated by a complex factor

$$\beta = |\beta| \exp(-i\theta). \quad (3.3)$$

Because of the high field in S_2 , the transit time there is negligible so that the hole current arriving at P_2 is

$$j(p, S_2)e^{i\omega t} = \beta je^{i\omega t}. \quad (3.4)$$

In addition to this current, there is a dielectric displacement current in S_2 which is converted to hole conduction current in P_2 . If the voltage drop across S_2 is

$$V(S_2, t) = V(S_2) + v(S_2)e^{i\omega t}, \quad (3.5)$$

then the ac displacement current is

$$j(D, S_2)e^{i\omega t} = i\omega C_2 v(S_2)e^{i\omega t}. \quad (3.6)$$

Now the total current is constant through the device, hence

$$j = j(p, S_2) + j(D, S_2) \quad (3.7)$$

which leads to

$$j = i\omega C_2 v(S_2)/(1 - \beta). \quad (3.8)$$

If $v(S_2)$ is substantially equal to the ac voltage across the unit, then the impedance is

$$\begin{aligned} Z &= v(S_2)/j = (1/i\omega C_2) + i\beta/\omega C_2 \\ &= (1/i\omega C_2) + (1/\omega C_2) |\beta| \exp i[(\pi/2) - \theta]. \end{aligned} \quad (3.9)$$

Evidently if $\theta > \pi$ and $\theta < 2\pi$, the second term will have a negative real part so that the diode will act as a power source.

If we neglect the ac electric field in N then β may be calculated in terms of the thickness $L = x_2 - x_1$ of the layer and the potential drop across the layer. This latter arises from the concentration ratio N_{d1}/N_{d2} between the two sides of N . Since the donor charge density is neutralized substantially entirely by electrons, and since almost no electron current flows, the electron concentration difference must result from a Boltzman factor (at $10^{17}/\text{cm}^3$ Fermi-Dirac statistics are not needed) and this leads to

$$\Delta V_n = (kT/q) \ln(N_{d1}/N_{d2}) \quad (3.10)$$

for the potential drop across N . In N the electric field is thus

$$E = \Delta V_n/L. \quad (3.11)$$

The differential equation for hole concentration for a disturbance of frequency ω is

$$\dot{p} = i\omega p = -\mu p E \frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2}. \quad (3.12)$$

This linear differential equation has two linearly independent solutions. These must satisfy at x_2 , the left edge of the space charge layer S_2 , the boundary condition that the hole density is practically zero.^{3,1} The appropriate solution is

$$p = e^{i\omega t} [e^{k_1(x-x_2)} - e^{k_2(x-x_2)}], \quad (3.13)$$

where

$$Lk_1 \equiv (x_2 - x_1)k_1 = \alpha[1 + (1 + i\gamma)^{1/2}], \quad (3.14)$$

$$Lk_2 \equiv (x_2 - x_1)k_2 = \alpha[1 - (1 + i\gamma)^{1/2}] \quad (3.15)$$

where

$$\alpha = q\Delta V/2kT, \quad (3.16)$$

$$\gamma = 4\omega D_p/u^2, \quad (3.17)$$

$$u = \mu_p E = \mu_p \Delta V/L. \quad (3.18)$$

The current is

$$j(p, x, t) = q(up - D_p \partial p / \partial x) \quad (3.19)$$

and the ratio of currents at x_1 and x_2 , which is β by definition, is

$$\begin{aligned} \beta &= j(p, x_2, t) / (j(p, x_1, t), \\ &= \frac{Lk_1 - Lk_2}{Kk_1 \exp(-Lk_2) - Kk_2 \exp(-Lk_1)} \\ &= \frac{2(1 + i\gamma)^{1/2} e^\alpha}{[1 + (1 + i\gamma)^{1/2}] \exp \alpha (1 + i\gamma)^{1/2} - [1 - (1 + i\gamma)^{1/2}] \exp - \alpha (1 + i\gamma)^{1/2}}. \end{aligned} \quad (3.20)$$

The phase lag in β must exceed 180° or π in order to give negative resistance. It can be seen that this phase factor must result from the first exponential in the denominator by the line of reasoning suggested below: The real part of the exponent is larger than the imaginary part. Hence the absolute ratio of the two exponentials is at least 2π . For this condition the second term in the denominator is negligible compared to the first. Hence the phase of (3.20) is determined largely by the first exponential. As a helpful approximation we may neglect the second term and write

$$\beta \doteq \frac{2(1 + i\gamma)^{1/2} \exp [\alpha - \alpha (1 + i\gamma)^{1/2}]}{1 + (1 + i\gamma)^{1/2}}. \quad (3.21)$$

Two limiting cases are worthy of special note:

$$(I) \quad \alpha \rightarrow 0, \text{ uniform } n\text{-layer}, \gamma \rightarrow \infty.$$

$$\beta \doteq 2 \exp -\alpha(i\gamma)^{1/2} = 2 \exp -(1 + i)(\omega/2D)^{1/2} L. \quad (3.22)$$

$$(II) \quad \alpha \rightarrow \infty, q\Delta V/kT \gg 1, \gamma \rightarrow 0.$$

$$\begin{aligned} \beta &= \exp [-i\alpha\gamma^{1/2} - \alpha\gamma^{1/2}/8], \\ &= \exp [-i(\omega L/u) - (DL/u)/(u/\omega)^2]. \end{aligned} \quad (3.23)$$

These expressions may be interpreted as follows: In case (I), flow is by diffusion and the propagation factors k_1 and k_2 take the form $\pm(\alpha/L)(i\gamma)^{1/2}$. For this case attenuation and delay terms in the exponential are equal, and the largest negative term occurs in Z when the phase angle is $5\pi/4$ (as may be verified by differentiation.) This leads to

$$\beta = 2(-1 + i)2^{-1/2} \exp(-5\pi/4) = 0.028(-1 + i), \quad (3.24)$$

which gives

$$Z = (1/\omega C_2)(-0.028 - i 1.028), \quad (3.25)$$

the impedance of a condenser with a negative Q of 37. In order to make an oscillator by coupling this to an inductance, an inductance with a Q of more than 37 must be used. It is obviously advantageous to reduce the magnitude of the negative Q .

For case (II) in its ideal form, the ac current simply drifts through the n -layer without attenuation. This produces a phase lag of ω times the transit time L/u . If this were the only effect involved, a capacitor with a negative Q of less than unity could be produced. In addition, however, there is attenuation due to spreading by diffusion. This effect is dependent upon the ratio of the spread by diffusion $(DL/u)^{1/2}$ to the separation of planes of equal phase in the drifting hole current. This latter separation is $2\pi u/\omega$. The square of this ratio appears in the attenuation term in the second form of β .

We shall estimate the effect of the attenuation term by taking

$$\alpha\gamma/2 = 3\pi/2, \quad (3.26)$$

so that the desired phase shift is obtained. The attenuation term is $\gamma/4$ smaller than this so that if $\gamma/4$ is considerably less than one, the attenuation in β will be small while the phase shift is correct. If we take $3\pi/2$ for the value of $\alpha\gamma/2$, then the value of γ becomes

$$\gamma = 4\omega D/u^2 = 6\pi kT/q\Delta V. \quad (3.27)$$

Thus the approximation on which (II) is based fails unless $q\Delta V/kT > 18$, a value which implies an enormous range of concentration in the n -layer. We must, therefore, investigate the case of gradients in the n -layer by more complete algebraic procedures.

We shall denote by $-\theta_1$ the phase shift in β due to the exponential in equation (3.21). The total phase shift θ is somewhat less since the algebraic expressions give a small positive phase shift of at most about 15° , which vanishes for large and small values of γ . Similarly the attenuation of β arises chiefly from the real part of the exponential since the absolute

value of the algebraic expressions lies between 2 for $\gamma = 0$ and 1 for $\gamma = \infty$.

It is instructive to express the real part of the exponent in terms of α and θ_1 . This is done as follows:

$$\alpha - \alpha(1 + i\gamma)^{1/2} = -\eta - i\theta_1. \quad (3.28)$$

This can be solved for η and $(1 + i\gamma)^{1/2}$:

$$\eta = (\alpha^2 + \theta_1^2)^{1/2} - \alpha, \quad (3.29)$$

$$(1 + i\gamma)^{1/2} = [(\alpha^2 + \theta_1^2)^{1/2} + i\theta_1]/\alpha. \quad (3.30)$$

From there it is seen that for a fixed value of θ_1 , the attenuation can be greatly reduced by increasing α . Unfortunately, this requires very large changes in concentration. For example with $\theta_1 = 3\pi/2$ and $\alpha = \theta_1$, the value of η is reduced to $\eta = 0.414 \theta_1$. However, the value of potential difference is

$$q\Delta V/kT = 3\pi, \quad (3.31)$$

giving

$$N_{d1}/N_{d2} \doteq 10^4. \quad (3.32)$$

For the case shown in Fig. 3.1, $\alpha = 1.15$ and for $\theta_1 = 5\pi/4 = 3.9$ we obtain

$$\eta = 2.95 = \ln_e 19.5 \quad (3.33)$$

This is an improvement of about 1 factor of e in the exponential compared to having $\Delta V = 0$. The value of β is

$$\beta = 0.082 \times \exp(-i 218^\circ), \quad (3.34)$$

and this leads to

$$Z = (\omega C_2)^{-1} (-0.05 - i 1.065). \quad (3.35)$$

Thus at the operating frequency, the diode appears to be a capacitor with a negative Q of 21.

Increasing the concentration change to a factor of 100, so that $\alpha = 2.3$, gives

$$\eta = 2.25, \quad (3.36)$$

$$\beta = 0.18 < -220^\circ, \quad (3.37)$$

$$Z = (\omega C_2)^{-1} (-0.116 - i 1.14), \quad (3.38)$$

$$Q = -10. \quad (3.39)$$

The calculations indicate that attenuation can be controlled to a considerable degree while maintaining the desired phase shift.

3.2. Justification of Consequences (1), (2) and (3)

In germanium at room temperature the product np is about 10^{27} under equilibrium conditions. At the first junction of Fig. 3.1 it is 10^{32} , implying a forward bias^{3,2} of (kT/q) 2.3×5 . In order to maintain this forward bias a flow of electrons must be furnished to N . There are several ways of accomplishing this. In the first place, the reverse bias across S_2 draws a reverse current of thermally generated electrons from P_2 . This current can be controlled by controlling the lifetime and temperature in the P_2 region. Alternatively, electrons may be injected into P_2 ; some of these will diffuse to S_2 and arrive at N . Still another means of controlling the bias across S_1 is to make contact to N itself. Since only the dc bias need be controlled, the series resistance across N itself is unimportant; the source should be of high impedance.

The decrease in density of 10^4 across the junction in carrier concentration implies a potential difference of 9.2 (kT/q) . Most of this potential difference occurs where the carrier concentration is negligible. Hence the space charge theory may be applied. Furthermore, the acceptor concentration is much higher than the donor concentration. Hence the space charge extends chiefly into the donor region and we may write^{3,3}

$$\Delta V_1 = (2\pi q N_d / \kappa) W^2 \quad (3.40)$$

for the relationship between width W of the space charge region and voltage drop ΔV .

If this voltage drop has an ac component, then a charging current will be required to change W . This current is determined by the admittance

$$\omega C = \omega \kappa / 4\pi W \quad (3.41)$$

of the space charge region.

At the same time injected hole and electron currents flow across the junction. The admittance associated with the hole current is approximately

$$A = (i\omega/D)^{1/2} \sigma_{p1}, \quad (3.42)$$

where σ_{p1} is the hole conductivity just inside the n-layer.^{3,4} Actually, as discussed below, the admittance is somewhat higher.

The ratio of the admittances is

$$\begin{aligned} \frac{|A|}{\omega C} &= \left[\frac{4\pi}{K} \frac{\omega \sigma_{p1}}{D} \cdot \frac{4\pi W^{-2}}{\kappa \omega^2} \right]^{1/2}, \\ &= \left[\frac{4\pi \sigma_{p1}}{K \omega} \cdot \frac{q \Delta V}{kT} \cdot \frac{\sigma_{p1}}{\mu_r q N_D} \right]^{1/2}. \end{aligned} \quad (3.43)$$

For our example this expression is much greater than 1 as may be seen as follows: The first fraction is the ratio of the dielectric decay constant to ω . This is 10^3 or more larger than ω need be. The next term is about 10 and the last term is the ratio of hole to electron density at x_1 and is about 10^{-2} . Hence the ratio of impedances is about 10:1.

We shall next consider why the expression for A for holes must be examined more closely. The admittance formula used above applies to the case of zero field to the right of the junction. The aiding field will increase the flow of holes into the n-layer and raise the admittance somewhat. Correcting for this will increase A in respect to ωC and will thus strengthen rather than weaken the argument.

Also in the expression for A , no account was taken of the transit time across the region W . If we assume a uniform field in this region for purposes of making estimates, then the solution of equation (3.20) may be applied. Since now u corresponds to drift velocity due to qkT/q of voltage drop across W which is much less than L in length, it is evident that γ will be less by a large factor in this region compared to its value in N . This leads to the conclusion that phase lags will be unimportant in this region.

We have neglected the effect of electron injection into P_1 . By the customary arguments for unsymmetrically doped junctions, it follows that this current is very small compared to the hole current.

This justifies consequence (1).

Consequence (2) may be justified as follows: At x_1 all the ac current $j \exp(i\omega t)$ is carried by holes. If a pure drift case occurred, the hole current might be reversed at some point in the n-layer and be $-j \exp(i\omega t)$. Under these conditions the electron current would have to be $2j \exp(i\omega t)$. Under no conditions, however, will the electron current be larger than this. This maximum possible electron current will require an electric field and this field will also affect the hole flow. Since the electron conductivity is at least 10 times larger than the hole conductivity, the hole current due to the ac field will only be about $1/10$ of j at most. Thus the hole current is only slightly affected by the ac field.

Consequence (3) follows from the fact that the reverse biased junction

S_2 has much higher impedance than S_1 . S_1 has higher impedance than that for hole injection into N . The impedance for hole injection into N corresponds to the hole conductivity in the n -layer over a distance comparable with the thickness of N . However, the impedance of N itself is that due to the much larger number of electrons in it and is thus much less than the impedance of S_1 . Thus it follows that the impedances across S_1 and across N are much less than across S_2 . This conclusion is not affected by the modification of impedance of S_2 due to hole flow across it.

3.3. Modifications

The treatment presented above has been based upon the conditions (a) to (e). Some of these are advantageous from the point of view of operation but others have been introduced to simplify the treatment. Among the latter is the condition that the current across S_1 is carried chiefly by holes. If the current were chiefly capacitative at this junction, then the voltage would lag 90° behind the current. This adds a desirable phase lag in the hole injection across S_1 and thus requires less phase shift in the n -layer. By suitably adjusting the ratio of capacitative and inductive admittances, a net improvement in Q may be obtained.

4. THE TRANSIENT RESPONSE IN A UNIPOLAR STRUCTURE

In the previous section the electric field produced by the injected holes had a negligible influence on the motion of the injected holes. In effect

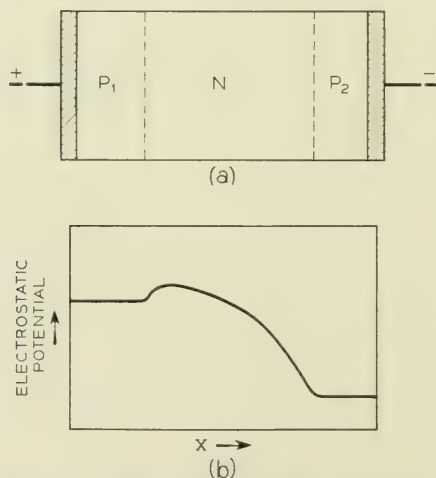


Fig. 4.1 — Space charge limited hole flow.

this was due to the bipolar nature of the mode of operation considered, the majority carriers in the region N acting to shield the minority carriers from their own space charge.

In this section we shall deal with unipolar diodes in which only one type of carrier is present in sufficient number to have a major effect. In these the influence of the space charge of the carriers upon their motion plays an important role.

Fig. 4.1 illustrates one example of the type of structure covered by the theory of this section. It is again a p-n-p structure like that considered in Section 3. However, in this case the dimensions, the donor density and the applied potential are such that the space charge "punches through" the device.^{4,1} Under these circumstances a condition of space charge limited emission is set up so that holes are injected from the positive region P_1 to just such an extent that their flow is limited by their own space charge. This limitation is associated with the maximum of potential just inside N .

The potential maximum is evidently a "hook" for electrons generated thermally in P_2 and in N . Under some circumstances electrons may accumulate and form a layer in which there is no electron flow and hole flow is carried equally by diffusion and drift. Such *stagnant* regions will tend to be suppressed if P_1 is made of short lifetime material, so that electrons are siphoned out of N , or if p at the maximum is larger than p for intrinsic material and the lifetime is locally low.

We shall treat the transient response of this structure of Fig. 4.1 from the point of view of the *impulsive impedance* discussed in Section 2. Accordingly we suppose that a steady current J flows per unit area. At $t = 0$ an added pulse of current occurs carrying a total charge of δQ_i per unit area, the subscript " i " signifying initial condition. Our problem is to determine how this added charge is carried by a transient disturbance in the hole flow and what is the resultant dependence of voltage upon time; by definition the added voltage across the device is

$$v(t) = \delta Q_i D(t). \quad (4.1)$$

Since we are dealing with a planar model, we shall suppose that the initial condition at $t = 0$ corresponds to added charges δQ_i and $-\delta Q_i$ on the metal plates on the P -regions. These charges set up an added field

$$\delta E_i = \delta Q_i / K, \quad (4.2)$$

where

$$K = \kappa \epsilon_0 \quad (4.3)$$

in MKS units. The initial value $v(0)$ is then simply δE_i times the total width of the structure.

The first effect, which takes place in a negligible time in respect to the frequencies involved, is the dielectric relaxation of the field in P_1 and P_2 . The added current due to δE_i leads to an exponential decay of δE in these regions with a transfer of δQ_i and $-\delta Q_i$ to the two boundaries of N . If P_1 and P_2 are thin compared to N , the resulting drop in $v(t)$ is small. In any event it can be shown by the reasoning at the end of Section 2 that this contribution to $D(t)$ adds simply the series resistance of P_1 and P_2 to the impedance.

The next effect is the transport of δQ_i on left side into N by hole flow over the potential maximum. It will be easier, however, to discuss this process after the treatment of the transient effects that occur in N itself. Consequently, we shall at this point assume that after a time, short compared with the important relaxation time in the structure, the disturbance of hole density is as shown in Fig. 4.2(a).

Fig. 4.2(a) shows added charges $+\delta Q_i$ and $-\delta Q_i$ produced by a disturbance denoted as δp in the hole density. The charge $-\delta Q_i$ on the right side is produced by an increased penetration of the space charge into P_2 ; it is similar to that produced by increasing reverse bias on a p-n junction.

Fig. 4.2(b) shows the corresponding disturbance in electric field. This disturbance is denoted by δE which is a function of x and t . Evidently

$$v(t) = \int_0^L \delta E(x, t) dx. \quad (4.4)$$

and this is the area under the δE curve.

The other parts of the figure indicate qualitatively a subsequent stage in the motion and decay of δp and δE . Our problem is to formulate mathematically this decay process. We shall treat the decay process in terms

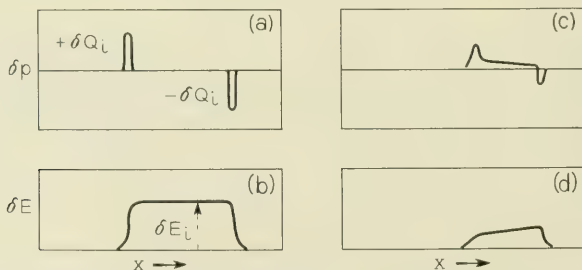


Fig. 4.2 — The initial stage and a subsequent stage of the transient.

of the effect of drift in the electric field and neglect the effects of diffusion. This procedure can be justified by the fact that as soon as a hole had reached a point where the potential has fallen by kT/q below the maximum, its flow is governed by drift rather than diffusion and the predominance of drift continues to increase towards the right.^{4,2}

If drift in the field is the predominant cause of hole flow, then the equations governing the situation in N are

$$J + \delta J = (\rho + \delta\rho)(u + \delta u), \quad (4.5)$$

where the terms with δ represent the transient effects and those without represent the steady state solution, $\rho = qp$ is the charge density of the holes and u their drift velocity. The equation for the change of E with distance is

$$K(\partial/\partial x)(E + \delta E) = \rho_f + \rho + \delta\rho, \quad (4.6)$$

where ρ_f is the fixed charge density due to donors and acceptors. (We neglect any effect of traps.) The steady state equation for E is thus

$$K(dE/dx) = \rho_f + J/u. \quad (4.7)$$

In a region where ρ_f is independent of x , this equation may be reduced to quadratures by writing

$$K dE/(\rho_f + J/u) = dx; \quad (4.8)$$

the left side is then a known function of E through the dependence of u upon E .

It is convenient to introduce a time-like variable s which is the transit time for the dc solution. Evidently

$$ds = dx/u = K dE/(\rho_f u + J). \quad (4.9)$$

For the case of space charge limited current, s may be conveniently measured from the potential maximum. Even though the solution is invalid at that point, the integrals converge and the contribution from the region within kT/q of the maximum is small.

We shall assume that the equations for the steady state case have been solved and that the functional relationships are known between E , x , v and s .

The differential equation for δE may then be obtained as follows: To the left of the pulse in δp in Fig. 4.2(a), δE is zero. From equation (4.6) we have

$$K\partial E(x)/\partial x = \delta\rho. \quad (4.10)$$

Integrating this from the region where E is zero gives

$$K\delta E(x, t) = \int_0^x \delta\rho(x, t) dx. \quad (4.11)$$

Equation (4.11) states that the dielectric displacement at x is equal to the excess charge between the potential maximum and x . Evidently during the transient following Fig. 4.2(a), the rate of change of this extra charge is $-\delta J(x, t)$ since the dc current is flowing in at the left and an excess current δJ flows out at the right. Hence we have

$$\begin{aligned} K\partial\delta E/\partial t &= -\delta J, \\ &= -(\delta\rho u + \rho\delta u). \end{aligned} \quad (4.12)$$

For the change in drift velocity we may write

$$\delta u = (du/dE) \delta E = \mu^* \delta E. \quad (4.13)$$

For high electric fields u increases less rapidly than linearly with E and μ^* is less than the low-field mobility.^{4,3} For very high fields μ^* is nearly zero and there are theoretical reasons for thinking that there may be a range in which μ^* is negative. We shall return to this point in the next section.

In Fig. 4.3 we show a diagrammatic representation of the transient so-

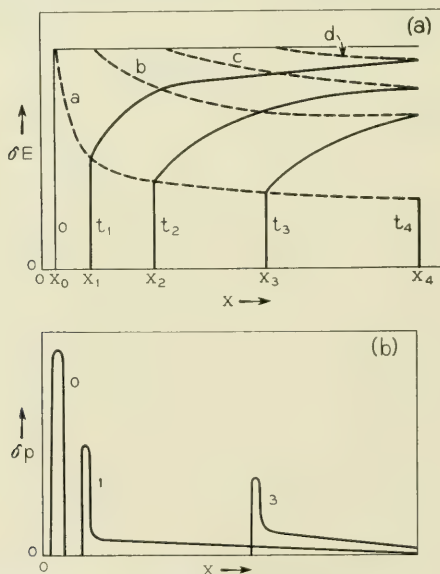


Fig. 4.3 — Graphical representation of the dependence of δE upon time.

lution. Each of the dashed lines represents the decay of δE as measured in a moving coordinate system: Thus we consider δE measured at a position $x(s_0 + t)$; this is a position that moves with the dc velocity u . This δE is evidently expressed in terms of $\delta E(x, t)$ by writing $x = x(s_0 + t)$:

$$\delta E \text{ in moving system} \equiv \delta E_m(s_0, t) = \delta E[x(s_0 + t), t]. \quad (4.14)$$

The differential equation for δE_m is

$$\begin{aligned} (\partial/\partial t) \delta E_m &= (\partial \delta E / \partial t)_x + (\partial \delta E / \partial x)_x \partial x / \partial t, \\ &= (\partial \delta E / \partial t)_x + (\partial \delta E / \partial x)_x u, \\ &= -(u \delta \rho + \rho \delta u) / K + (\delta \rho / K) u, \\ &= -(\rho \mu^* / K) \delta E = -\nu \delta E, \end{aligned} \quad (4.15)$$

where the quantity

$$\nu \equiv \rho \mu^* / m \quad (4.16)$$

is an effective dielectric relaxation constant being the *differential conductivity* $\rho \mu^*$ divided by the permittivity K .

Evidently ν is a function of position x only and may be expressed as $\nu(s)$ through the dependence of x upon s . Thus we may write

$$(\partial/\partial t) \delta E_m(s_0, t) = -\nu(s_0 + t) \delta E_m(s_0, t) \quad (4.17)$$

which has a solution

$$\delta E_m(s_0, t) = \delta E_m(s_0, 0) \exp [-g(s_0 + t) + g(s_0)], \quad (4.18)$$

where

$$g(s_0 + t) = \int_{s'}^{s_0 + t} \nu(s) ds. \quad (4.19)$$

The lower limit s' is chosen for convenience so as to avoid singularities in $g(s)$. This integration shows that δE_m decays exponentially as the electrical field would decay in a material whose dielectric relaxation constant changed with time just as ν changes as observed on the moving plane.

Fig. 4.3 shows on the dashed lines the decay of δE_m on the moving planes. Since δE_m is zero to the left of the initial pulse in Fig. 4.2(a), it remains zero on all moving planes which follow the pulse of δQ_0 . This justifies the statement made earlier. The solid curves labelled t_1, t_2 etc. show the spatial dependence of δE for times t_1, t_2 , etc. after the charge δQ_1 is added.

The values of the transient voltage $v(t)$ at time t_1 , for example, is the

integral under the curve t_1 . This curve is zero for $x < x(t_1)$ and for $x > x(t_1)$ it is

$$\delta E(x, t_1) = (\delta Q_i/K) \exp [-g(s_0 + t_1) + g(s_0)], \quad (4.20)$$

where

$$x = x(s_0 + t). \quad (4.21)$$

If the total transit time across N is S so that

$$x(S) = L, \quad (4.22)$$

then

$$v(t_1) = \int_{x(t_1)}^L \delta E(x, t_1) dx. \quad (4.23)$$

From this expression we can derive the desired formula for $D(t)$. For this purpose the integral over dx is replaced by an integral over s . At time t the range of s is evidently from t to S and $dx = u(s) ds$. From this we obtain:

$$\begin{aligned} D(t) &= v(t)/\delta Q_i, \\ &= (1/K) \int_t^S \exp [-g(s) + g(s-t)] v(s) ds. \end{aligned} \quad (4.24)$$

From Fig. 4.3 we can see that there are competing tendencies in the decay of $D(t)$ some of which tend to produce the desired convex shape discussed in Section 2 and others the concave shape. The effect of the dielectric relaxation constant is adverse and tends to produce an exponential decay. On the other hand the advance of the pulse of holes from left to right in Fig. 4.2 proceeds in an accelerated fashion with the result that the range of x over which δE is not zero decreases at an accelerated rate. If the dielectric relaxation were zero, this would result in the desired convex upwards shape.

The resultant shape of the $D(t)$ curve is thus sensitive to the exact relationship of the transit time and dielectric relaxation. This can be illustrated by giving the results of analysis for a p-n-p structure, neglecting diffusion and considering μ to be constant. The solutions of the dc equations are readily obtained for this case and have been published.^{4,4} For convenience we repeat them here:

$$E = (J/\mu\rho_f)(e^{\alpha s} - 1), \quad (4.25)$$

$$x(s) = \mu JK(\mu\rho_f)^{-2} (e^{\alpha t} - \alpha t - 1), \quad (4.26)$$

$$L = x(s) = (JK/\mu\rho_f^2)(e^\beta - \beta - 1), \quad (4.27)$$

$$\beta \equiv \alpha s.$$

From these it is found that

$$\ln g(s) = (1 - e^{-\alpha s}). \quad (4.29)$$

This leads to

$$\begin{aligned} D(t) &= (J/\mu\rho_f^2) [e^\beta + e^{\alpha t} (\alpha t - \beta - 1)] \\ &\equiv (J/\mu\rho_f^2) D(\beta, \alpha t). \end{aligned} \quad (4.30)$$

For $t = 0$ this reduces correctly to L/K .

Figure 4.4 shows the resulting shape of the D curves with β as a parameter. Large values of β correspond to cases in which the hole charge density is small compared to ρ_f and to relatively long relaxation constants. For them the desired convex upward shape results.

Figure 4.5(a) and 4.5(b) show the real and imaginary parts of the impedance expressed in terms of $Z(\beta, \theta)$:

$$Z(\omega) = \int_0^T e^{-i\omega t} D(t) dt \quad (4.31)$$

$$= (KJ/\mu^2\rho_f^3) Z(\beta, \theta),$$

$$\theta = \omega T. \quad (4.32)$$

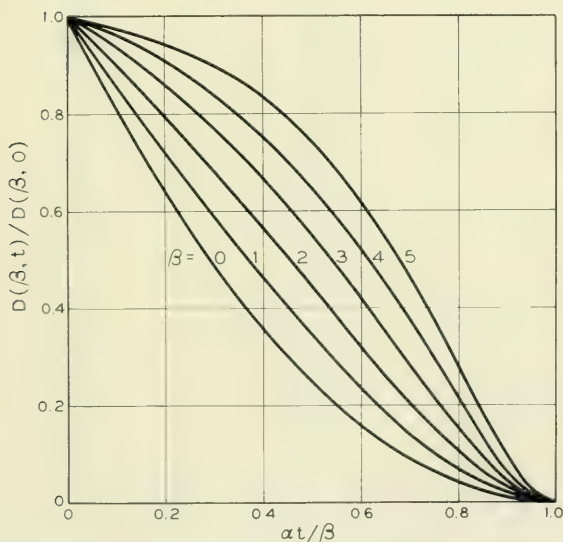


Fig. 4.4 — Impulsive impedance for various values of β in p-n-p structure.

It is seen that values of $-Q$ as small as about 10 can be obtained for $\beta \geq 3$.

In the next section we shall consider modifications which may result from variations in μ^* and for changes in geometry.

We must return to the question of how the charge $+\delta Q_i$ passes the potential maximum. In order that the theory given above apply, it is necessary that the time required for δQ_i to enter the drift region be short compared to the transit time. At the potential maximum the charge density may be estimated by the methods previously dealt with in the theory of space charge limited emission. Initially $+\delta Q_i$ appears to the left of the maximum and the field at the maximum is δE_i . This field will

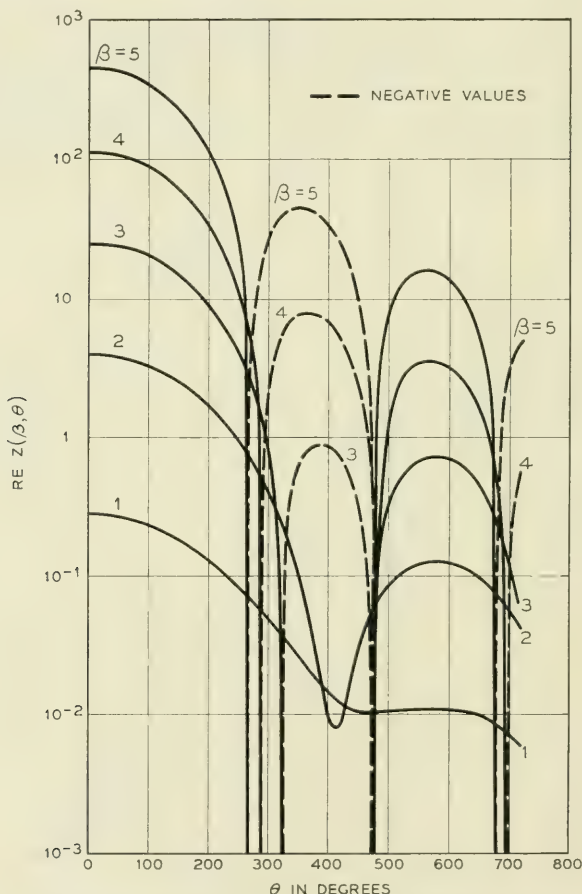


Fig. 4.5 — Impedance of a p-n-p structure. (c) Real part of impedance.

then relax with a relaxation constant of about $\mu\rho(\text{max})/K$ where $\rho(\text{max})$ is the hole charge density. Actually the relaxation may be somewhat quicker because the concentration gradient of the added holes also contributes to the flow over the maximum. Since the charge density at kT/q below the maximum is comparable to that at the maximum the entire relaxation process will proceed at about this rate. Thus a criterion for the applicability of the theory is that $K/\mu\rho(\text{max})$ be much less than S , the transit time or total decay time for $D(t)$.

5. MOBILITY AND GEOMETRY EFFECTS

5.1. *The Effect of a Region of Negative μ^**

In very high electric fields holes may be expected on the basis of theory to exhibit a negative value of μ^* . This theory^{5,1} is founded on the idea that a hole can lose energy to phonons at a certain maximum aver-



Fig. 4.5 — Impedance of a p-n-p structure. (b) Imaginary part of impedance.

age rate $P_{\max}$. ($P_{\max}$ is the *staukonstante* of Krömer.) The holes probably achieve this rate when their energy is near the middle of the valence band. Under these conditions the power input from the electric field must be no greater than $P_{\max}$:

$$qEu \leq P_{\max} \quad (5.1)$$

From this it follows that

$$u \leq P_{\max}/qE, \quad (5.2)$$

so that the drift velocity will decrease with increasing field at sufficiently high fields.

Furthermore, if the width of the valence band is less than the energy gap, then a hole cannot acquire enough energy to produce hole electron pairs. Thus in such a case, the negative resistance range should be reached before breakdown effects occur.

In Fig. 5.1 we illustrate the general trends of the u versus E curve, to be expected if the *stau эффект* occurs. As is indicated, the maximum drift velocity will be referred to as u_m . It occurs at a field E_m . Since we are here concerned with principles rather than details, no attempt has been made to indicate the square root range in which u is proportional to $E^{1/2}$. This range has been observed by E. J. Ryder^{5.2} and shown by G. C. Dacey^{5.3} to control hole flow in space charge limited hole currents in germanium and has been treated theoretically.^{5.4} Dacey^{5.5} has also investigated the effect of the square root law upon the $D(t)$ curves for the p-n-p structure of Section 4 and reports that the effects are so unfavorable that no negative resistance is to be expected.

The *stau эффект* opens the attractive possibility of making negative resistance devices in which the current decreases with increased dc voltage so that negative resistance will be exhibited over a wide frequency range. Unfortunately, when the boundary conditions are taken

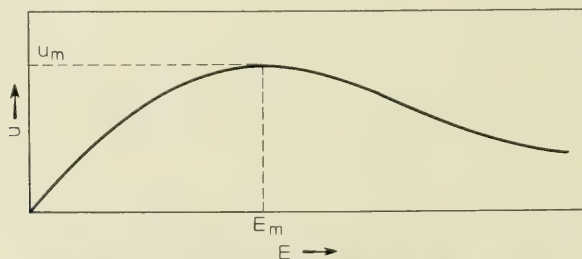


Fig. 5.1 — Qualitative representation of drift velocity versus field as affected by “*stau эффект*.”

into account, it is found that a device in which most of the current flow occurs in a negative μ^* region does not necessarily show a dc negative resistance characteristic. On the other hand, such a structure may have a very favorable $D(l)$ characteristic.

We shall illustrate these conclusions by considering a $(p^+)p(p^+)$ structure having heavily doped ends, so that ohmic non-injecting contacts may be made to the ends. Fig. 5.2 shows the potential distribution and hole distribution for two cases of applied potential. The first case, represented in (a) and (b), corresponds to moderate fields such that the peak velocity u_m is not reached.

In the second case the voltage is so high that the average value of $E = V/L$ exceeds the critical field E_m . Under these conditions u is approximately equal to u_m over a large part of the P -region and a substantial portion of the voltage drop occurs near one end. An increase of the applied voltage occurs chiefly at this end with a small increase in current. Thus no negative dc resistance occurs.

The above conclusions are reached by considering the differential equation for the space charge again as in Section 4 neglecting diffusion and starting with $E = 0$ at the left edge of the P -region. This leads to

$$dx = K dE / [(J/u) - \rho_a], \quad (5.3)$$

where we have introduced

$$\rho_a = -\rho_f = -qN_a \quad (5.4)$$

for the charge density of the acceptors. Evidently if J exceeds J_m where

$$J_m = \rho_a u_m, \quad (5.5)$$

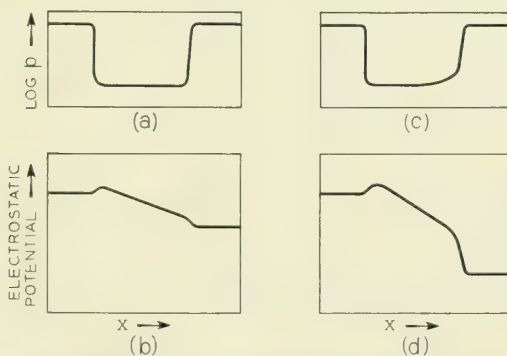


Fig. 5.2 — Hole density and potential distribution including influence of “staueffekt.”

the denominator is everywhere positive and x increases monotonically with E . If J is only a little larger than J_m , there will be a large region of x in which E is nearly equal to E_m . Outside of this region, E increases much more rapidly with x . The relative scale of these distances may be estimated as follows: Suppose J is only a little larger than J_m , and consider the situation where E is about twice E_m so that u is about one half of u_m . Then

$$dx/dE \doteq K/\rho_a. \quad (5.6)$$

Under these conditions an increase of E by an additional E_m will require a distance

$$\Delta x \doteq KE_m/\rho_a. \quad (5.7)$$

If this value is much smaller than L , then the situation represented in Fig. 5.2(c) and (d) will occur. On the other hand if L is smaller than Δx , the region of space charge and high field will extend throughout most of the structure.

In any event equation (5.4) leads to positive resistance. This can be seen from the fact that increasing J always means a decrease in x for the same value of E and hence an increase in E at all values of x and thus an increase in voltage at any fixed value of x .

The above conclusion that a positive dc resistance will be exhibited by a structure like that discussed above may also be reached by considering the transient response. The theory of Section 4 may be once at be applied to this case by simply taking account of the fact that ν is negative for part of the structure and thus that δE_m increases with increasing s .

In Fig. 5.3 we illustrate a structure to which these considerations may be relatively simply applied, at least in a limiting case. It consists of four layers, the two outer being p^+ as before. Space-charge limited emission then enters the intrinsic layer which is of such a width that at its right hand boundary the electric field has a value E_3 that exceeds E_m . At this point the hole space charge is

$$\rho_3 = J/u_3 \quad (5.8)$$

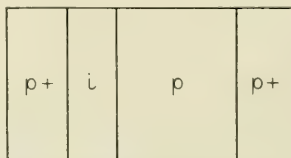


Fig. 5.3 — A structure having a region of uniform negative differential conductance.

where u_3 is $u(E_3)$. In the P -region this space charge is compensated by acceptors to produce a region of uniform field in which μ^* is negative.

If the P -region is wide compared to the I -region, then the transit time through it will also be relatively large. As a consequence δQ will be transferred quickly into the P -region. From that time on δE_m curves, like those of Fig. 4.3, will show an exponential increase with time and also with distance since for this case of constant u in the P -region, time and distance are linearly related. This will lead to a $D(t)$ of the form

$$D(t) = (u_3/K)(S - t) \exp |\mu^* \rho_3/K| t, \quad (5.9)$$

where the absolute value signs emphasize that for this case of negative μ^* there is a build-up in time. This form of D is always convex upwards and, in fact, if

$$S |\mu^* \rho_3/K| > 1, \quad (5.10)$$

it starts with a positive slope at $s = 0$ so that the transient voltage actually builds up initially with time.

Even an initially growing $D(t)$ does not give a negative resistance at low frequencies, however. As shown in Section 2, the dc resistance is simply the integral under the $D(t)$ curve and thus will still have a positive value.

5.2. Convergent Geometry

It is possible to obtain marked improvement of the $D(t)$ curves without the aid of the negative values of μ^* . This possibility is based upon convergent geometry. A possible case is illustrated in Fig. 5.4. In this case it is supposed that the field in the inner P -region is so large that a substantial reduction in μ^* has occurred. As a consequence, the decay of field in this region is relatively slow. Furthermore, since both the dc and

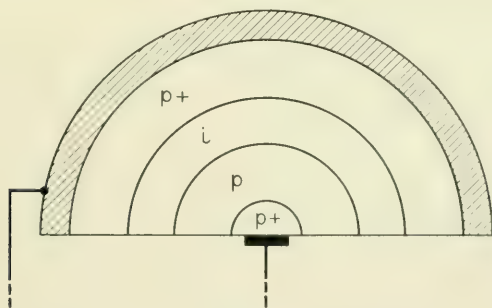


Fig. 5.4 — A convergent flow structure.

transient fields are high in this region, essentially because of the inverse square law, the principal contribution to $D(t)$ comes from this region. These two factors — relatively slow dielectric relaxation near the center and principal contribution to $D(t)$ from near the center — combine to give a $D(t)$ characteristic which holds up well until the pulse of injected holes reaches the inner region. This may result in a favorable convex upwards $D(t)$ characteristic.

ACKNOWLEDGEMENTS

The writer is indebted to a number of his colleagues for helpful discussions and to W. van Roosbroeck for the calculations for Fig. 4.4, and to R. C. Prim for Fig. 4.5.

REFERENCES

- 1.1. W. Shockley and W. P. Mason, J. of Appl. Phys., **25**, No. 5, p. 677, 1954.
- 3.1. W. Shockley, *Electrons and Holes in Semiconductors*, D. van Nostrand, N. Y., 1950, p. 312.
- 3.2. See Reference 3.1.
- 3.3. W. Shockley, B. S. T. J., **28**, p. 435, 1949, Section 2.4.
- 3.4. See References 3.1 or 3.3.
- 4.1. W. Shockley and R. C. Prim, Phys. Rev., **90**, pp. 753-758, 1953. See also G. C. Dacey, Phys. Rev., **90**, pp. 759-763, 1953, and W. Shockley, Proc. I.R.E., **40**, pp. 1289-1314, 1952.
- 4.2. W. Shockley and R. C. Prim, Reference 4.1.
- 4.3. E. J. Ryder, Phys. Rev., **90**, pp. 766-769, 1953, references.
- 4.4. See Shockley and Prim, Reference 4.1.
- 5.1. This theory of mobility in higher fields has been published by H. Krömer, Zeits f. Physik **134**, pp. 435-450, 1953. Krömer considers the theory in connection with the values of α in point contact transistors but does not explore it as a power source. The present writer derived the same result in a more primitive form in 1948 as a potential means of obtaining high frequency power. However, experimental results by E. J. Ryder did not show evidence of this effect. The effect may possibly have occurred without being recognized because of the absence of an adequate appreciation of the importance of the boundary conditions discussed in this section.
- 5.2. See Reference 4.3.
- 5.3. G. C. Dacey, Phys. Rev., **90**, pp. 759-763, 1953.
- 5.4. W. Shockley, B. S. T. J., **30**, pp. 990-1043, 1951.
- 5.5. Personal communication.

Transistors and Junction Diodes in Telephone Power Plants

By F. H. CHASE, B. H. HAMILTON and D. H. SMITH

(Manuscript received November 30, 1953)

This paper describes the use of junction diodes, reference voltage diodes, and junction transistors in regulated rectifiers for telephone power plants. It discusses the pertinent characteristics of these semiconductor devices, together with illustrative circuits in which they are used to control the flow of direct current power.

1. INTRODUCTION

Recent articles in the literature have treated the theory and properties of semiconductor devices. In particular, papers by Messrs. Shockley, Ryder, Wallace and others have emphasized the theoretical aspects of the new devices; their reliability, reproducibility and performance at high frequencies to name only a few.^{1, 4, 5, 6, 7} In addition many papers have been published concerning their applications in the transmission and computer fields. There is also a field of application for these devices in the conversion and control of power, and this paper discusses some of these power applications.

1.1. Scope

The first three groups of sections in this discussion review the pertinent characteristics and practical engineering aspects of junction rectifier diodes (Section 2.1), reference voltage diodes (Section 2.2) and junction transistors (Section 2.3). The second three groups of sections concern respectively shunt transistor regulators (Section 3.1), series transistor regulators, (Section 3.2) and power regulating circuits employing magnetic amplifiers in combination with transistors and junction diodes (Section 3.3). The last two groups of sections treat specific applications (Sections 4.1 through 4.3).

2. DEVICE CHARACTERISTICS

2.1. *Junction Rectifiers*

A junction rectifier is made from a wafer cut from a single crystal of semiconductor material. The materials now being used for this purpose are germanium and silicon, but to date the use of germanium is more common than silicon. Pure germanium in its undisturbed or intrinsic state is a poor conductor; but its conductivity can be increased by disturbances such as cosmic rays, photons of light, external potentials, or by the addition of very small amounts of selected impurities. We are concerned here only with the addition of impurities. There are two classes of these impurities, called "donors" and "acceptors." The physical mechanism by which pure germanium becomes conductive depends on which of these two classes of impurities are present. Donor impurities result in a surplus of free electrons which can conduct current by negative charges passing through the germanium crystal. Thus the addition of donor impurities to pure germanium creates "n" type material. Presence of acceptor impurities results in a shortage of electrons creating "holes," which have positive charges. These holes are mobile and they can conduct current through the crystal.⁷ Thus the addition of acceptor impurities to pure germanium creates "p" material.

When an abrupt change is made from p to n type material inside the crystal a rectifying junction exists at the boundary between the two materials. This p-n junction exhibits rectifier action in that it will conduct current very easily from p toward n; but, in its rectifier operating range, only minute currents can be made to flow from n toward p. We say that this junction has a low forward resistance and a high reverse resistance. All rectifiers have these characteristics to a greater or lesser degree and the p-n junction rectifier characteristics have been compared elsewhere to other rectifier devices.²

There are two methods of producing the junction inside the crystal. It can be obtained by growing part of the crystal from p type material and part from n type. This is called a "grown" junction. It can also be obtained by diffusing impurities into the crystal after it has been grown. This has been called an "alloy" process, a "fused junction" process, or a "diffused junction" process.

2.11. *Junction Rectifier Terminology*

Before discussing the characteristics of junction diodes, it may be helpful for the reader to consider the terminology employed. As in other

rectifying cells, there are two directions of current flow, forward and reverse. Each diode has a positive and a negative terminal, and we define the positive terminal as that terminal towards which forward current flows *within the diode*. Likewise, the negative terminal is that terminal towards which reverse current flows *within the diode*. In Fig. 1(a), terminal 1 is the negative terminal and terminal 2 the positive. The circuit convention for the diode is a shorthand method of indicating the polarity of the diode to the engineer. If a battery is connected to a diode as shown in Fig. 1(b), forward current will flow, and if connected per Fig. 1(c), reverse current will flow. If the battery is replaced by a source of alternating current, forward current will flow through the diode during the half cycle that terminal 1 is positive, and reverse current will flow during the half cycle that terminal 2 is positive. The rectifier is said to “conduct” during the first half cycle and to “block” during the second half cycle, for the resistance in the conducting direction is very much less than the resistance in the blocking direction.

The figure of merit of a diode is a measure of this ease of conduction and the effectiveness of the blocking action. The ease of conduction can readily be determined on a static basis by applying a dc voltage to the diode as shown in Fig. 1(b) and plotting forward current through the diode as a function of applied voltage. Likewise, the blocking characteristic can be determined if a circuit per Fig. 1(c) is employed.

2.12. Typical Junction Rectifiers

Fig. 2 is a photograph of several sizes of typical junction diodes. The diodes shown have a range of forward current from several milliamperes (Diode I) to hundreds of amperes (Diode IV). Diode I is made from a crystal of silicon and the balance are made from germanium. Most rectifying diodes have a particular field of use dictated mainly by their power handling capacity in the forward direction of current flow, although Diode I is of interest because of its unusual reverse or blocking characteristic, as will be pointed out later in this paper.

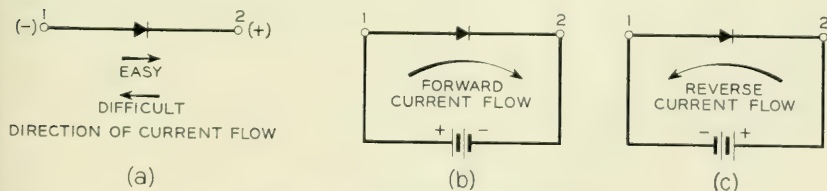


Fig. 1 — Rectifier terminology.

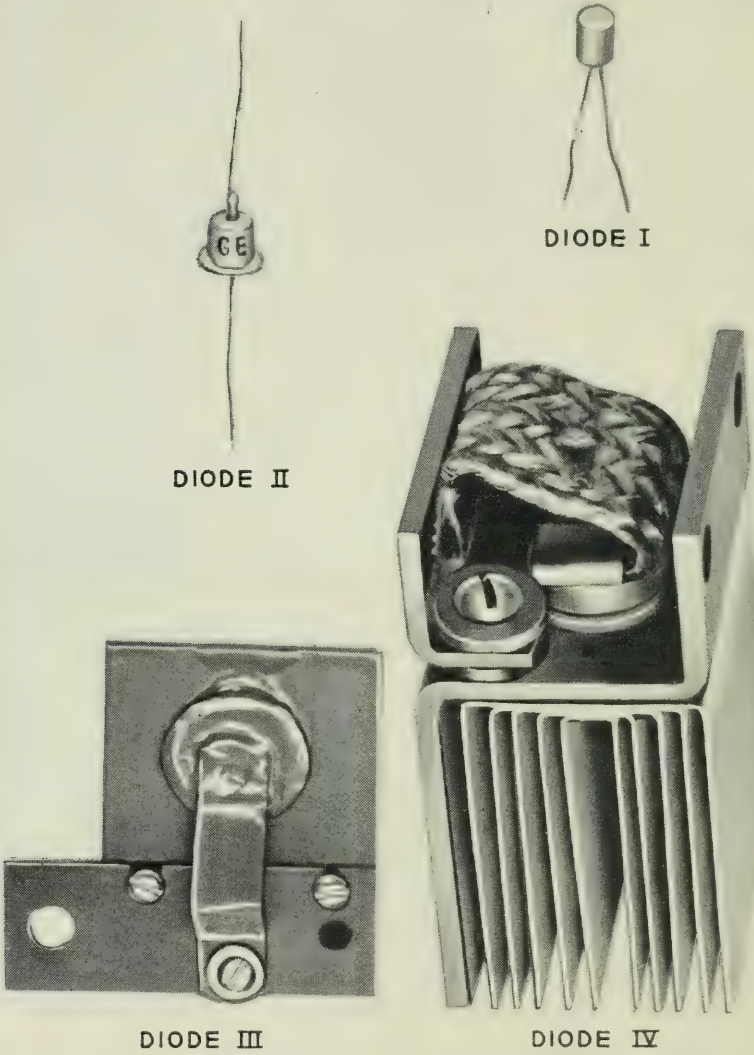


Fig. 2 — Typical junction rectifiers. Diodes II, III and IV, courtesy of the General Electric Company.

Fig. 3 is a plot of the static forward and reverse characteristics of the four diodes shown in Fig. 2. The characteristics were obtained using the circuits in Figs. 1(b) and 1(c), respectively, measurements being made in still air at room temperature. The curves in the first quadrant, $(+E + I)$ are the forward characteristics and the curves in the third quadrant $(-E - I)$ are the reverse characteristics. Notice that the scales are different in these quadrants. In general, at any other temperature the curves would shift their positions with respect to the reference axes. This must be taken into account by the circuit designer.

2.13. Junction Temperature

We will limit further discussion of general characteristics to those of Diode IV, for in many respects this is the most interesting rectifier for

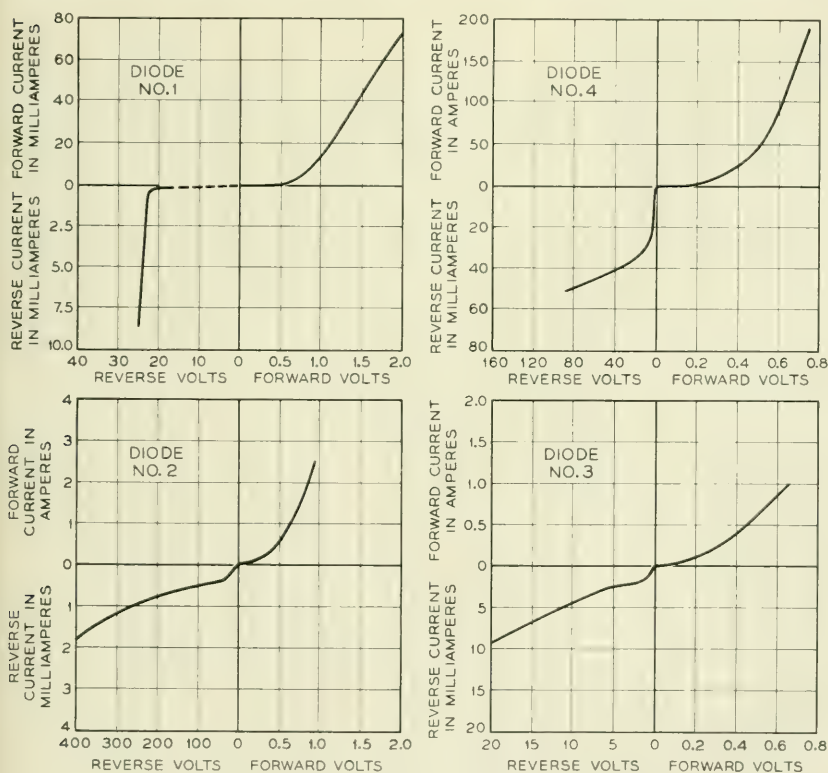


Fig. 3 — Junction rectifier static characteristics.

power applications. Laboratory experience indicates that it is not desirable to operate the junction of this diode above 65° to 70° centigrade. The value of this critical temperature is not accurately known on account of the difficulty in measuring the junction temperature inside of the crystal. However, below the critical temperature, those changes in characteristics which are associated with changes in junction temperature are reversible, that is, if the temperature is raised and then reduced, the characteristics will shift back to values previously experienced at the reduced temperature. Beyond the critical junction temperature any change in the reverse characteristics is permanent and has the effect of reducing the reverse resistance. In an operating circuit, this effect leads to progressively greater permanent damage to the diode. Lowered reverse resistance allows more reverse current to flow, increases the reverse power dissipation and elevates the temperature of the junction causing further reduction of the reverse resistance, and so on until the diode no longer blocks.

Thermal damage to the junction can be prevented by removing heat. This method is employed with the diode under discussion by forcing air through the cooling fins at a high velocity. The quantity of air needed depends on the amount of heat generated in the junction, the efficiency of the cooling fins and the temperature of the air employed for cooling. In most Bell System applications, the maximum temperature of the

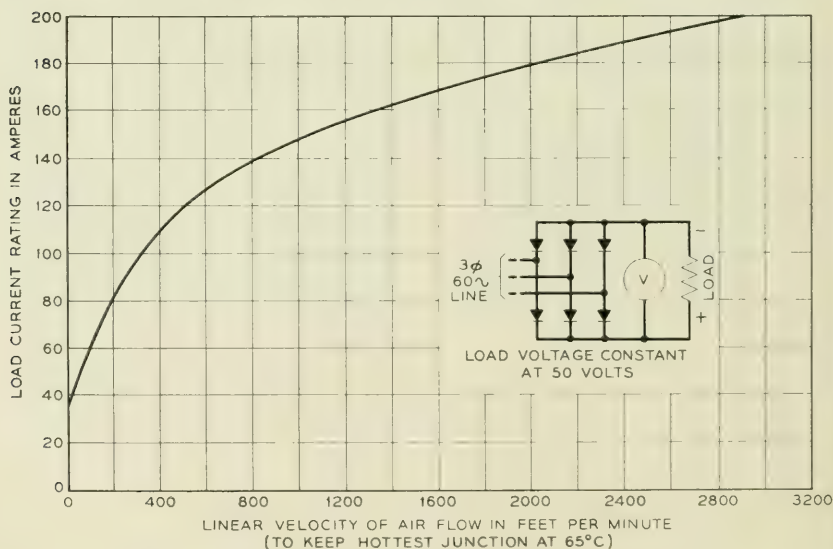


Fig. 4 — Junction rectifier forced cooling characteristics.

ambient air is 40°C , which permits the junction temperature to be 25 to 30°C above the air temperature before the critical value is reached.

A typical load-current versus air velocity curve is shown in Fig. 4. The curve is based on a 65°C junction temperature measured by thermocouples attached to the radiating structure near the junction and 40°C ambient air. Notice that the curve is taken with a working circuit composed of six diodes in a three-phase full wave bridge arrangement. In general, engineers developing rectifier circuits find that curves showing the properties of combinations of rectifying diodes are more useful than single diode characteristics, except where the properties of the diode are such as to make it useful as a valve, or as a reference standard, as is the case of Diode I in Fig. 2. This leads directly to a more detailed consideration of the blocking or reverse characteristics of junction rectifiers.

2.2. *Reference Voltage Diodes*

2.21. *General*

In the case of silicon junction diodes it has been possible to reduce the reverse current to a very low value for reverse voltages up to a value called the "saturation voltage." When the saturation voltage is reached the electrons and/or holes which comprise the leakage current are given sufficient energy to create other electron-hole pairs which add to the original reverse current. This process is cumulative and leads to large increases in current for small further increases in voltage. The effect is illustrated by the reverse voltage-current characteristic for Diode I in Fig. 3. This curve shows the reverse current to be quite low for voltages less than 22 volts. This portion of the characteristic is called the "high resistance region." As voltage is further increased the curve goes through a "transition region" to the "saturation voltage region" at 23 volts where voltage is nearly constant over a wide range of current. The voltage saturation characteristic makes the diode suitable for use as a source of reference potential in the control of power. Those readers who wish to study the basis of these properties will find the theory covered elsewhere in the literature.^{3,8}

The rectifier selected for study in this Section is Diode I. This is a p-n junction rectifier made from silicon. It has been constructed to obtain a reasonably constant saturation voltage as shown in Fig. 5. In order to show the wide range of current values where this voltage is substantially constant, Fig. 5 is plotted to a logarithmic scale. In this connection it is interesting to note that the saturation voltage can be controlled in manufacture from a few volts to several hundred volts. This

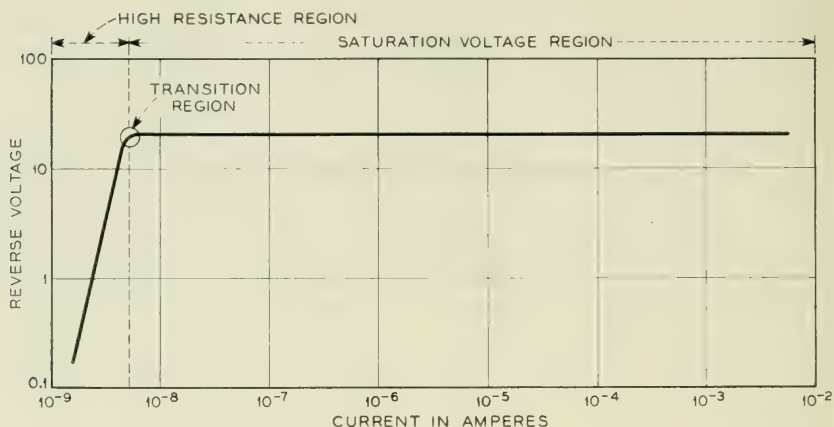


Fig. 5 — Reverse characteristics of a reference voltage diode.

range can be compared to the 60 to 150 volts range of cold cathode voltage regulator tubes which are also used as sources of reference potential.

2.22. Saturation Voltage Utilized in Regulating Circuits

In all check back (feedback) regulating circuits the potential to be regulated is compared to a reference potential. This comparison is a form of subtracting the two values so that the changes in the potential to be regulated produce a large percentage change in the difference or error voltage. The methods by which this is accomplished in direct current circuits are illustrated in Section 3. A stable source of reference potential is required for this type of regulation. When the saturation voltage of a silicon junction diode is used for this purpose, we have called the device a "reference voltage diode."

2.23. Effect of Temperature on Saturation Voltage

In order to evaluate the stability of Diode I in its saturation voltage region a small section of Fig. 5 has been redrawn in Fig. 6 using a linear scale. Additional curves are included in Fig. 6 to show the change of voltage with ambient temperature variations. The slope of the 30 degree curve in Fig. 6 is equivalent to a resistance of 200 ohms in series with a 23-volt battery with current flowing through this combination from an external source. The change of potential with ambient temperature is equivalent to a 0.07 per cent change per degree C. It should not be inferred that these are limiting values, for diodes have been tested which exhibit slopes of less than 10 ohms and temperature coefficients of less

than 0.01 per cent per degree C. The specific applications covered later in this discussion show methods to compensate for slope and temperature variation when necessary.

2.3. Junction Transistor Action

2.31. Two-Rectifier Analysis

In junction transistors there are two p-n junction rectifiers contained in the semiconductor material. Of the materials now in use germanium is the more prevalent. Remembering the results of adding donor and acceptor impurities to obtain n and p type materials covered in section 2.1 these two rectifiers are obtained by interposing a layer of p type material between two layers of n type making an *n-p-n transistor* or interposing a layer of n type material between two layers of p type making a *p-n-p transistor*. The electrical connections are designated as the collector terminal, the emitter terminal and the base terminal. Both types of transistors (n-p-n and p-n-p) have a rectifying junction between the collector and base terminals and another rectifying junction between the emitter and base terminals. The polarity of the collector and emitter rectifying junctions determines whether the transistor is n-p-n or p-n-p.

Figs. 7(a) and 7(b) are simplified diagrams illustrating respectively the internal circuits of n-p-n and p-n-p transistors. The figures show the characterizations of transistors by means of a two-rectifier analogy. Although a transistor may be somewhat over-simplified by this method of characterization, the analogy permits the power engineer to approximate the operation of transistors in familiar terms. Experience in the development of the circuits described later in this article has proven that the analogy is valid under circumstances where the operation of the transistor as a dc amplifier is of interest.

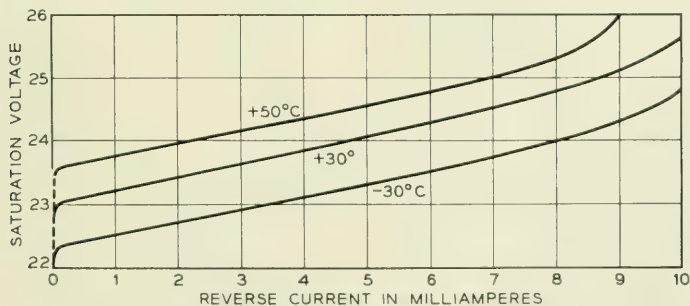


Fig. 6 — Saturation voltage characteristics of a reference voltage diode.

In an n-p-n transistor the collector and emitter terminals are the positive electrodes of the rectifiers, see Fig. 7(a), and in a p-n-p transistor the collector and emitter terminals are the negative electrodes of the rectifiers, see Fig. 7(b). The base terminal is the common point of the two rectifiers. In a given transistor each rectifier has a saturation voltage, usually stated in the characteristics, which must not be exceeded in normal operation. Thus, the saturation voltage of the collector rectifier determines the maximum instantaneous collector potential. The emitter rectifier also has a saturation voltage which determines the maximum potential which can be applied between the base and the emitter. The saturation voltage of the collector rectifier usually differs from the saturation voltage of the emitter rectifier.

2.32. Transistor Action

If a source of potential, E_{ce} in Figs. 7(a) and 7(b) is connected between the collector and emitter terminals, the resulting current will flow in series through the collector rectifier in its reverse direction and through the emitter rectifier in its forward direction. This is the direction of current flow for transistor action to take place. In Fig. 7 the reverse resistances of the collector rectifiers are shown and the forward resistances of the emitter rectifiers are also shown.

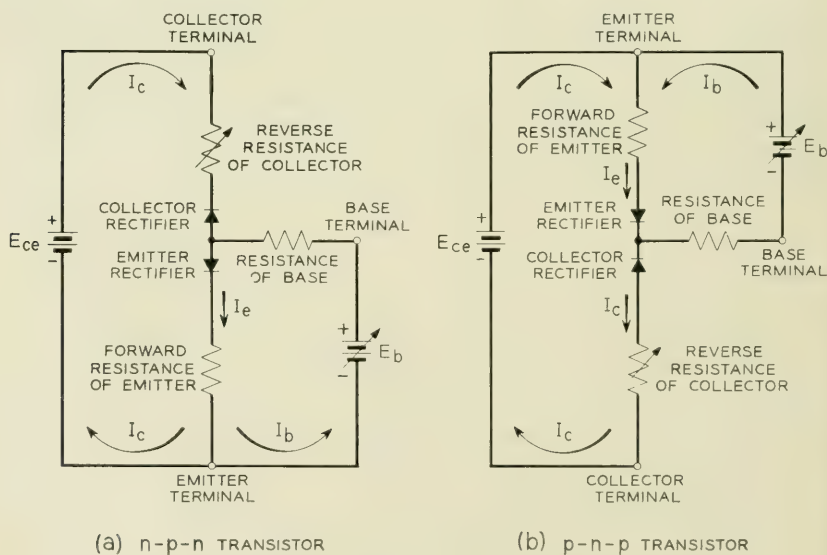


Fig. 7 — Junction transistor analogy.

2.33. Current Gain

Now when a second relatively small potential is connected between the base and emitter rectifier (E_b in the sketches) additional current, I_e will flow through the emitter rectifier in the forward direction and I_c will also increase. This increase in I_c caused by the increase in I_e is transistor action. The increase in I_c is related to the increase in I_e by the factor alpha (α) as written below:

$$\Delta I_c = \alpha \Delta I_e. \quad (1)$$

The application of Kirchhoff's current law to the sketches in Fig. 7 gives the change in I_b as follows

$$\Delta I_b = \Delta I_e - \Delta I_c. \quad (2)$$

By combining equations (1) and (2), ΔI_c can be written as a function of ΔI_b only

$$\Delta I_c = \frac{(\alpha)}{(1 - \alpha)} \Delta I_b. \quad (3)$$

The usual value of α for junction transistors is near but slightly less than unity. In a typical case α might be 0.98. This value when substituted in equation (3) shows the current gain of the transistor, $\Delta I_c / \Delta I_b$ to be 49. Most of the circuits discussed in this paper are based on equation (3).

It has been shown how a small change in base to emitter potential with a small change in base current effects a large increase in collector current at a higher voltage. This explains how large power gains, of the order of 60 db, can be obtained from the junction transistor.

The sketches in Fig. 7 do not show why this transistor action takes place. The reasons for it involve the use of such solid state physics terms as the migration of electrons and holes through a crystal lattice, and the interposition of junction barriers. The "why" for transistor action is very important in the manufacture of transistors, and it has been thoroughly covered in the literature.^{1, 7} For present purposes it is only necessary to examine the static characteristics of an n-p-n transistor as shown in Fig. 8. This figure presents transistor characteristics in a manner which simplifies the explanation of the operation of the transistor control circuits covered later in this paper.

Referring to the curves in Fig. 8, it will be seen that in the straight portion of the 1.5-volt curve, a change of 50 microamperes in the base current will result in a change of about 2 milliamperes in the collector current. This illustrates the current amplification of transistors and the

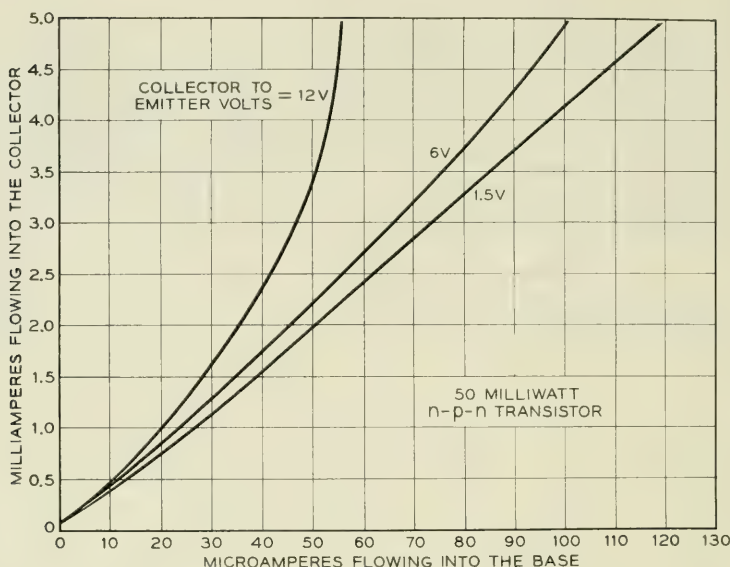


Fig. 8 — Junction transistor static characteristics.

current gain of this transistor is equal to 40. The measured α for this transistor was 0.976. Substituting in the current gain formula, equation (3) above, the calculated current gain is 40.6 which agrees with Fig. 8 within the accuracy of the measurements.

2.34. Low Voltage Characteristics

Again referring to the curves in Fig. 8, it will be seen that transistors operate at low collector to emitter potentials. The 1.5-volt curve is not the minimum potential at which this transistor will operate. Some transistors have good current amplification at potentials as low as two-tenths of a volt. When the base current is reversed, the characteristics in Fig. 8 can be extended to smaller collector current values. One might assume that the collector current can be reduced to zero by causing enough current to flow out of the base. This is not true. There is a minimum collector current, called the saturation current, and increasing current flow out of the base will not decrease the collector current below this value. This saturation current is assigned the symbol I_{C0} . This I_{C0} current is usually a few microamperes but it increases at the rate of 7 or 8 per cent per degree Centigrade increase in temperature of the collector junction. Transistors also have a critical junction temperature

which should not be exceeded under any operating conditions, and this must be kept in mind during the design of the regulating circuits.

2.35. Equivalent Circuit of a Transistor

Ryder and Kircher⁴ have shown that it is possible to convert the sketches shown in Fig. 7 into a small signal equivalent circuit using α and the three characteristic resistances of the transistor. These resistances are the emitter resistance r_e , the base resistance r_b and the collector resistance r_c . Two forms of equivalent circuit are shown in Figs. 9(a) and 9(b). In the equivalent circuit in Fig. 9(a) the active portion of the transistor is characterized as a current generator. This equivalent circuit is more directly related to the physical processes occurring inside the transistor than the equivalent circuit in Fig. 9(b) which characterizes the active portion of the transistor as a voltage generator. Although both equivalent circuits are useful the one in Fig. 9(a) is preferred in power work because r_c is much larger than the load resistance in many cases and can be neglected. Typical values for the equivalent circuit parameters are given in the caption of Fig. 9. The use of the equivalent circuits are further discussed in some of the articles listed at the end of this paper. The article⁵ by R. L. Wallace Jr. and W. J. Pietenpol is of particular interest in this connection.

2.36. Typical Junction Transistors

Fig. 10 is a photograph of two Bell System junction transistors made from germanium. The smaller one will dissipate 50 milliwatts, and the larger one is an exploratory model that will dissipate 2 watts when it is attached to a suitable heat sink. These transistors are hermetically sealed to protect them from the infiltration of moisture. The characteristics shown in Fig. 8 were measured using the smaller unit.

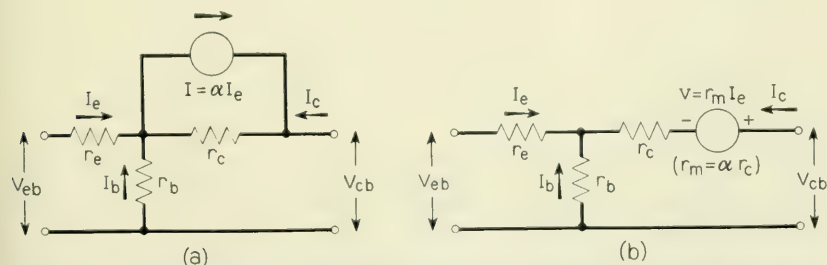


Fig. 9 — Junction transistor equivalent circuits. Typical values for a 50-milliwatt transistor: r_e , 25 ohms; r_b , 500 ohms; r_c , 5 megohms; α , 0.98; and r_m , 4.9 megohms.



Fig. 10 — Typical junction transistors.

Thus, note that there are two kinds of transistors with respect to the polarity of the electrodes. The n-p-n transistor operates with positive collector potential and the p-n-p requires negative potential on the collector. Both will amplify current changes in the base circuit into much larger current changes in the collector circuit. The transistors have similar equivalent circuits and parameters but all of their operating potentials and currents are reversed. It is also significant that the normal direction of current flow is out of the base terminal of the p-n-p transistor and that the normal direction of current flow is into the collector terminal of the n-p-n transistor. Likewise, the normal direction of current flow is out of the collector terminal of the p-n-p transistor and into the base terminal of the n-p-n transistor. This relationship between direction of current flow in n-p-n and p-n-p transistors is called reversed or complementary symmetry, and enables the circuit designer to cascade direct coupled transistors, alternating n-p-n and p-n-p. This is not possible with vacuum tubes because there is no tube that will operate with negative plate potential. It will be shown how this complementary symmetry can be used to advantage in multistage direct current amplifier circuits.

3. TYPICAL REGULATING CIRCUITS

3.1. *Shunt Regulators*

3.11. *Simple Diode Regulator*

If a load is connected to a source of power, the current through the load and thus the voltage drop across the load will depend on the potential of the source of power, the internal impedance of the power supply and the load impedance. The voltage drop across the load can be made very nearly independent of these three parameters by employing a circuit known as a shunt regulator.

A shunt regulator is a variable current device, connected in parallel with the load. Both the load and the shunt regulator draw current from the source of power through a common impedance. The operating requirement for a good shunt regulator is that the voltage drop across it must remain constant over a certain range of current. Certain types of cold cathode tubes such as the VR-150-30(0D3) exhibit this effect. It has been determined that certain semiconductor junction diodes exhibit the same effect. Note that diode No. 1 that is shown in Fig. 3 has a reverse voltage drop of about 24 volts over a range of reverse current from less than 1 milliamperere to almost 10 milliamperes. If such a device is connected in parallel with a load as in Fig. 11, shunt regulating action will take place. Consider the operation of the circuit in Fig. 11, first assuming that the load impedance is constant and that the potential of the power source increases. Additional current tends to flow from the source, but since the potential of the reference voltage diode designated "s" in Fig. 11 is fixed, this additional current develops an increased voltage drop across the regulating resistor, and the load voltage does not change. Similar reasoning can be applied to the case of a decrease in source voltage. Next assume constant source voltage and an increase or decrease in load resistance. This would normally tend to cause a change in the voltage drop across the load, but the shunt element draws respectively more or less current than normal and the load voltage again does not change.

The value of the regulated load potential in Fig. 11 is controlled by the saturation voltage of the diode and it cannot be adjusted to any other value. The accuracy of regulation is controlled by the slope of the reverse characteristics shown in Fig. 6. An additional limitation is that the usefulness of this type of regulator is controlled by the power handling capacity of the diode. The next section shows how these limitations can be circumvented by the addition of transistors.

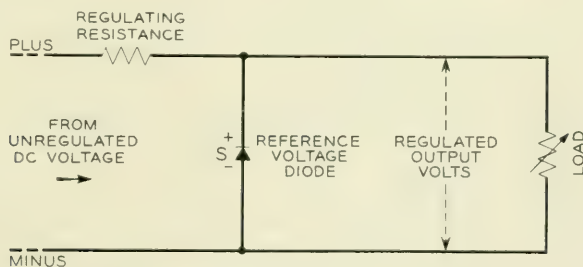


Fig. 11 — Simple shunt regulator.

3.12. Transistor-Diode Regular

To examine the operation of shunt transistor regulators consider the circuit shown in Fig. 12. In this, the transistor is shown by its standard convention where the upper slanting line represents the collector rectifier shown in Fig. 7 and the lower slanting line with the arrow on it represents the emitter rectifier. The direction of the arrow shows that, in this transistor, current flows out of the emitter, so it is an n-p-n transistor. The (c), (b) and (e) designations also help to locate the collector, base and emitter. Notice that the emitter current, shown by the arrow I_e , flows through the reference voltage diode in its reverse direction. The rectifier symbol with an adjacent "s" is a convention for this diode.

In Fig. 12 a portion of the load voltage is applied to the base of the transistor by means of the adjustable potentiometer. The potential of the emitter is held constant with respect to the negative output potential by the saturation voltage of the diode. The base-to-emitter voltage is thus equal to a proportion of the load voltage minus the saturation voltage. The potentiometer is adjusted so that the base potential is slightly positive with respect to the emitter when the desired value of voltage appears across the load. This value of load voltage is called the regulated voltage. Current I_b then flows into the base, current I_c flows through the regulating resistance, and $I_b + I_c$ combine to form I_e . Now assume that the regulated voltage (E) increases by an amount ΔE . The base voltage becomes more positive with respect to the negative terminal by the proportion of ΔE developed across points 1 and 2 of the potentiometer. Since the emitter potential is held constant by diode "s" and cannot change, the net effect is to increase the base to emitter potential. This change in base to emitter potential causes an increase in collector current,

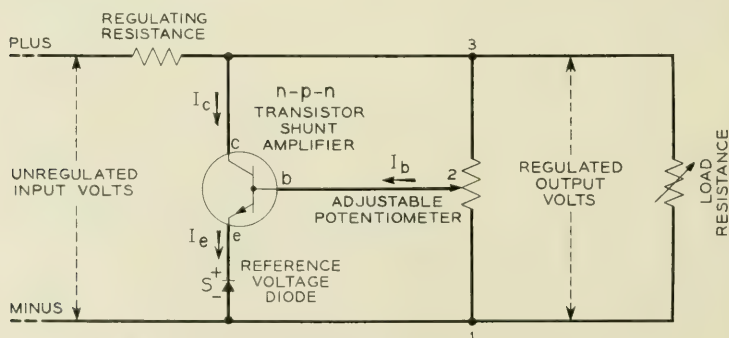


Fig. 12 — Transistor shunt regulator using one transistor.

a consequent increase in voltage drop in the regulating resistor, and a decrease in the load voltage. The correcting process continues until the load voltage returns to the regulated value, and takes only a small fraction of a second. The process is essentially the same for a decrease in load voltage, except that the base to emitter potential decreases, the collector current decreases, the voltage drop across the regulating resistor decreases and the load voltage rises to the regulated value.

The value of the regulated output voltage is determined by the adjustment of the potentiometer. Of course in a practical shunt regulator circuit, the adjustable range of the potentiometer would have to be limited to correspond with the operating range of the transistor. The maximum allowable positive potential between the base and the emitter is limited by the safe value of the maximum collector current. The maximum allowable negative potential between the base and the emitter is limited by the saturation voltage of the emitter rectifier.

The accuracy of this shunt regulator circuit is restricted by the slope of the characteristic curves for the reference voltage diode. All of the changes in base and collector currents required for regulation flow through this diode and cause changes in the saturation voltage. The addition of the transistor does not increase the accuracy of regulation but only allows adjustment of the regulated output potential to a value which is greater than the standard potential. However additional stages of transistor current amplification minimize the reference potential changes by restricting the range of current excursions through the diode. An example of a multistage shunt regulating circuit is given in Fig. 13.

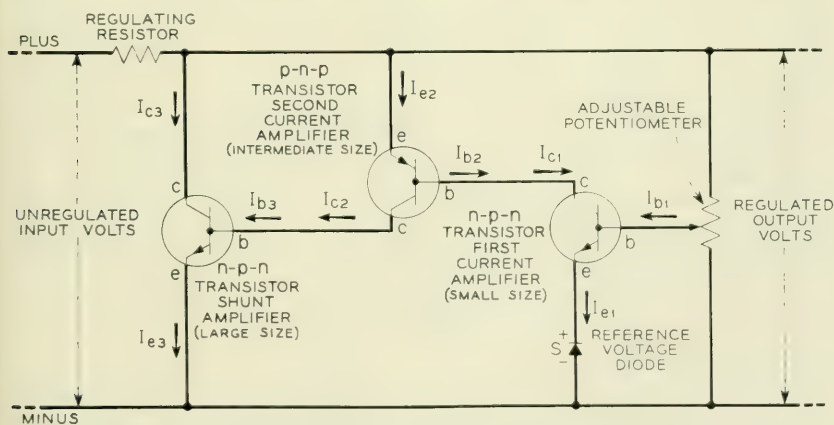


Fig. 13 — Transistor shunt regulator using three transistors.

3.13. *Multistage Transistor Shunt Regulator*

In Fig. 13 two additional transistors have been added to the simple shunt regulator of Fig. 12 in order to increase the accuracy of regulation. The first stage (subscript 1 is used for the transistor currents in this stage) compares the output potential to the reference voltage, drives the second stage (subscript 2) which in turn drives the third stage (subscript 3). The first stage transistor operates in a similar fashion to the transistor in Fig. 12, except that its collector current now is the base current of the second transistor. The collector current of the second transistor is the base current of the third transistor. The second and third stage transistors amplify the collector current of the first transistor. The shunt regulating current is the sum of the currents in all three transistors. An examination of Fig. 13 will reveal that the first transistor is an n-p-n, the second transistor is a p-n-p, and the third transistor is an n-p-n and that no coupling networks are used. This illustrates the advantages of complementary symmetry.

In Fig. 13 different sizes are specified for the three transistors. The transistor shown for the first current amplifier might be a 50-milliwatt transistor operating at a collector potential of about 10 volts. Then the maximum base current of the second stage p-n-p transistor should not exceed 5 milliamperes and, with an assumed current amplification of 20 times, the maximum collector current of the second stage could be 100 milliamperes. Such a transistor has been developed. With 100 milliamperes flowing into the base of the large n-p-n transistor and an assumed current amplification of 20 times the maximum shunt regulator current would be about 2 amperes which would compensate for considerable load current variations. Large size transistors such as would be necessary in the third stage are now under exploratory development within the industry.⁶

The circuit in Fig. 12 can be modified to use a p-n-p transistor and several other modifications can be made. Similar modifications can be made in the circuit shown in Fig. 13. It is not within the scope of this article, however, to show all the permutations and combinations of transistor regulator circuits that are usable. Section 3.2 below covers some typical transistor series regulator circuits.

3.2. *Series Regulators*

Precise voltage control can be obtained with shunt regulators but series regulator circuits are usually more efficient. This comes about because the shunt regulator wastes the shunt current plus the voltage

drop across the regulating resistance whereas the series regulator wastes only the voltage drop across the series device. At light load the power dissipated in the shunt current is usually greater than the power dissipated in the series circuit. With a transistor used as the series regulator device this difference in efficiency is more pronounced because of the small collector voltage that can be used for the full load current. This collector voltage is the voltage drop across the series transistor as shown in Fig. 14.

3.21. Simple Series Regulator

Fig. 14 shows a simple transistor series regulator circuit. A p-n-p transistor is shown connected so that all the load current must pass through it. The comparison of the output voltage to the reference potential in the current amplifier of Fig. 14 is accomplished by holding the emitter at a constant potential with respect to the *positive* output terminal. Note the difference between this method and that covered in the previous section on the shunt regulator 3.12, where the emitter was held at a constant potential with respect to the *negative* output terminal. Now, when the output potential increases by an amount ΔE , the base voltage becomes more *negative* with respect to the positive terminal by the proportion of ΔE developed across points 2 and 3 of the potentiometer. Since the emitter cannot change with respect to the point of reference (the positive terminal), the net effect is to *decrease* the base to emitter potential and the collector current for an increase in output voltage. The collector current decrease is amplified by the current gain of the p-n-p series transistor to decrease the load current, reducing the output

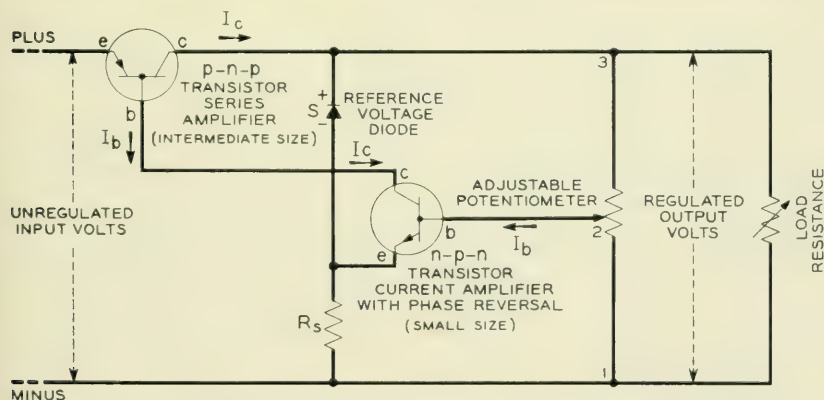


Fig. 14 — Transistor series regulator constant voltage regulation.

voltage, and thus regulating it. The value of the regulated output voltage is again determined by the adjustment of the potentiometer. The ohmic value of the R_s resistor in Fig. 14 is selected to keep the current flowing through the reference voltage diode in its saturation voltage region.

Fig. 14 is the simplest form of a transistor series regulator circuit. It requires two transistors whereas the most simple form of a transistor shunt regulator shown (Fig. 12) requires only one transistor. But the added current gain of the second transistor in Fig. 14 results in better regulation than can be obtained with Fig. 12. If desired the circuit in Fig. 14 can be modified to change the series transistor to the negative output lead by using the complementary p-n-p first current amplifier and an n-p-n series transistor. This illustrates another advantage of the complementary symmetry of the two types of transistors. Also, if more gain is required, additional transistor stages can be used employing the principles outlined above.

3.22. Series Current Regulator

The circuits covered so far regulate for constant output voltage. Similar transistor regulator circuits can be developed which will regulate for constant output current. One of these is shown in Fig. 15. In this circuit the load current produces a voltage drop across the regulating resistance and, in the n-p-n transistor, this voltage drop is compared to the reference voltage. The difference between these two potentials controls the n-p-n transistor base current and this base current is amplified by the current gain of both transistors to control the load current.

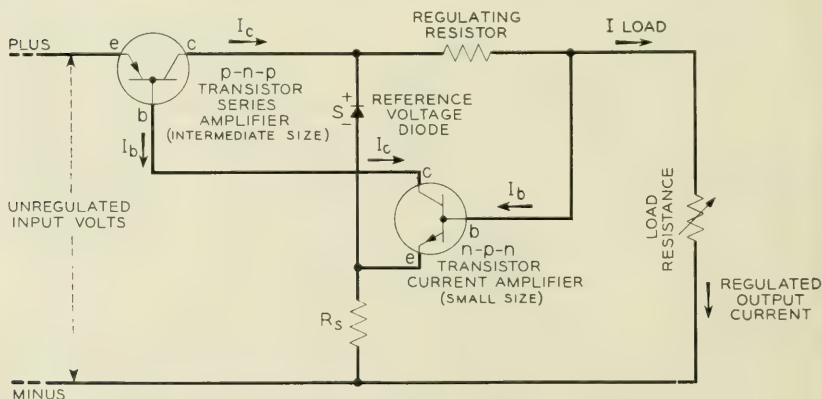


Fig. 15 — Transistor series regulator constant current regulation.

This circuit is phased so that the load current will be increased when it is too small and decreased when it is too large. The values of the regulating resistor and the reference voltage determine the value of the regulated load current. Additional current amplifier stages can be included or the circuit can be modified to change the series transistor to the minus lead as covered above.

3.3. Transistors Combined With Magnetic Amplifiers

3.31. General

Transistors can be used to control directly the flow of power to a load as pointed out in the sections on series and shunt regulators. However, their direct use is limited to moderate voltages (below 100 volts) or moderate currents (up to 1 ampere) with transistors now contemplated.

3.32. Transistors as DC Preamplifiers

In cases where regulation of higher power is required, it is expedient to combine transistor circuits with other devices having higher power-handling capacity. One type of combination is shown in Fig. 16, where a transistor is used to amplify weak dc error signals to a magnitude sufficient for driving a magnetic power amplifier.

In Fig. 16, emitter (*e*) of the n-p-n transistor is held at a fixed negative voltage with respect to the positive output of the power supply by the reference voltage diode ("S"). Another negative voltage derived from the output voltage of the power supply through potentiometer (*P*) is applied

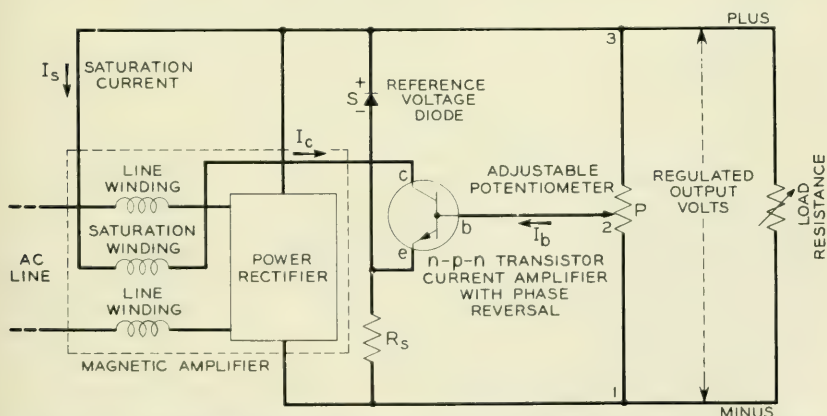


Fig. 16 — Transistor control circuit for a magnetic amplifier regulated rectifier with constant voltage regulation.

to base (b) of the transistor. This latter voltage is made a little smaller than the emitter voltage so that the base (b) is positive with respect to the emitter. Now assume that the load voltage increases for some reason such as an increase in the line voltage or a decrease in the load current. A portion of the increased load voltage appears across points 2 and 3 of potentiometer (P), and tends to make the base voltage more negative. Since the base is slightly positive with respect to the emitter, the net effect of making the base more negative is to decrease the base-to-emitter voltage. Through transistor action, the collector current, which is also the saturation current of magnetic amplifier, decreases and the ac impedance of the line windings rises. The line windings absorb more input voltage and the output voltage is brought back very nearly to the original value before the change.

The circuit of Fig. 16 is of interest because it can control larger amounts of power than can be handled by transistors alone and, in addition, it is capable of faster regulating action than an all-magnetic regulating circuit with the same loop gain. The use of the transistor in this circuit eliminates the need for one or more stages of milliwatt-size magnetic preamplifiers.

3.33. Increased Gain in Voltage Regulators

Additional amplification to improve the regulation can be added to Fig. 16 in two ways. Several stages of transistor current amplification can be added or more magnetic amplifier stages can be used. Of course

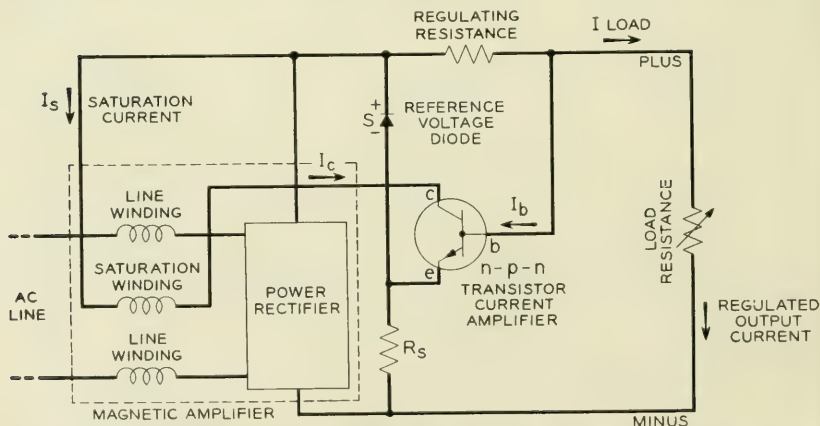


Fig. 17 — Transistor control circuit for a magnetic amplifier regulated rectifier with constant current regulation.

a combination of the two methods is also feasible. Additional magnetic amplifiers have the disadvantage of adding time delay. Transistor action likewise is not instantaneous because it takes a finite amount of time to move the charge over a finite distance in the crystal lattice. However transistor action is much faster than the time required to change the current in practical magnetic amplifiers.

3.34. Current Regulators

Fig. 17 shows a simple transistor control circuit to obtain constant current regulation with a magnetic amplifier regulated rectifier. The operation of this circuit is similar to Fig. 15 and its description will not be repeated.

3.35. Temperature Effects

One limitation of the foregoing transistor regulating circuits is the sensitivity of collector current to ambient temperature variations. The collector current increases with increasing temperature even if the base-to-emitter bias is held constant. This is the result of three factors. (1) I_{c0} , the uncontrolled portion of I_c increases greatly as covered in Section 2.34; (2) the emitter resistance (r_e) decreases causing I_b to increase, and (3) α changes. The effect of the temperature sensitivity of the collector can be greatly reduced by using a differential or push-pull circuit of the type illustrated in Fig. 18.

3.36. "Push-Pull" DC Amplifier

The push-pull circuit uses two emitter-coupled n-p-n transistors and is in many respects similar to a cathode-coupled vacuum tube amplifier.

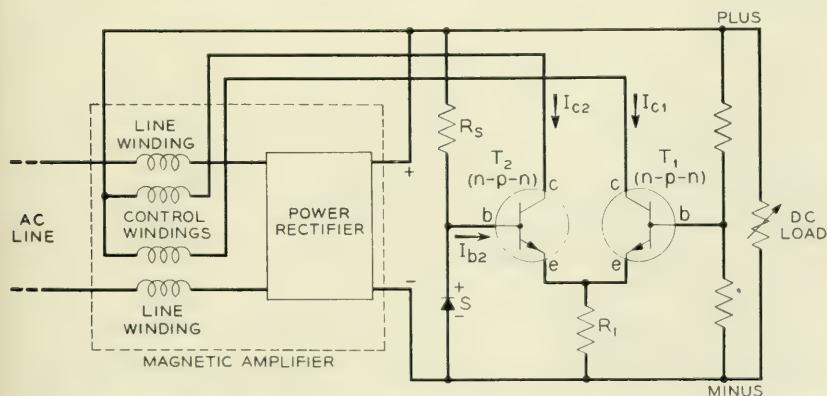


Fig. 18 — Push-pull transistor control circuit for a magnetic amplifier regulated rectifier with constant voltage regulation.

A voltage proportional to the regulated output is connected to the base of transistor ($T1$). Fixed resistors are shown in the base circuit of ($T1$) in Fig. 18, but a variable potentiometer could be used. The reference voltage diode "s" applies a constant reference voltage to the base of transistor ($T2$). If the output voltage tends to increase, more collector and emitter current flows in transistor ($T1$) due to the increase in its base-to-emitter voltage. This increase in emitter current of transistor ($T1$) flows through resistor ($R1$) and tends to raise the emitter voltage of transistor ($T2$). Since the base potential of transistor ($T2$) is fixed, the effect is to decrease the base-to-emitter voltage of ($T2$) and its collector and emitter currents decrease. The result is an increase in I_{c1} and an almost equal decrease in I_{c2} . If the two saturation windings on the magnetic amplifier are oppositely poled, the changes in I_{c1} and I_{c2} represent a net decrease in the control ampere turn input to the magnetic amplifier. As before, the magnetic amplifier responds by absorbing more voltage. If, however, I_{c1} and I_{c2} both increase equally due to an increase in ambient temperature, no net change is made in the control ampere turn input to the magnetic amplifier. Thus if the two transistors are perfectly matched, and the reference voltage diode has a low temperature coefficient, temperature changes will have little effect on the output regulated voltages.

A further advantage is the reduced variations in the current through the reference voltage diode. As in the case of the other circuits additional stages of transistor or magnetic amplification can be added to increase the loop gain and the precision of regulation.

4. APPLICATIONS

4.1. *General*

The last sections of this discussion cover some specific applications of the principles discussed above. Section 4.21 covers a one-stage transistor shunt regulated rectifier as a grid battery eliminator for phase controlled thyatron tube rectifiers. Section 4.22 covers a transistor voltage amplifier circuit as a grid battery eliminator for magnitude controlled thyatron tube rectifiers. A two-volt, three-ampere regulated rectifier covered in Section 4.31 illustrates how a low voltage, high current, regulated rectifier with a transistor and magnetic amplifier control circuit can be obtained. Section 4.32 covers a 65-volt, 200-ampere regulated rectifier for telephone central office battery charging. It uses the p-n junction rectifier devices covered in Section 2.1 and a modification

of the transistor control circuits for magnetic amplifier regulation covered in Section 3.36.

4.2. Grid Battery Eliminators

4.21. Phase Controlled Thyatron Tube Rectifiers

In thyatron tube regulated rectifiers the standard potential for the checkback regulator is often obtained from dry cells. The annual replacement of dry cell batteries is an appreciable maintenance expense, particularly in those cases where the rectifiers are installed in isolated or unattended locations. This section covers a transistor shunt regulated rectifier as a substitute for the dry batteries. Its circuit is illustrated in Fig. 19.

The circuit in Fig. 19 is the same as Fig. 12 with the addition of the compounding resistor and the thermistor. The compounding resistor is added to compensate for the slope of the reference voltage diode in its saturation voltage region (see Fig. 6). The thermistor is added to compensate for ambient temperature variations of this diode and the transistor.

The compounding resistor adds ac line voltage compounding. The transistor base current regulating signal in Fig. 19 is increased by the compounding resistance whenever the ac voltage is increased. By selecting the proper ohmic value of the compounding resistor, the circuit in Fig. 19 can be arranged so it will deliver constant output voltage into a constant resistance load when the ac voltage is varied from 85 per cent

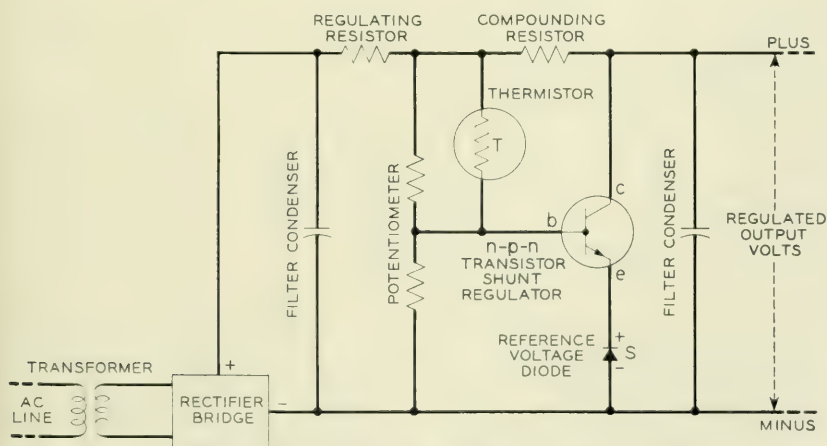


Fig. 19 — Grid-battery eliminator for phase-controlled thyatron rectifier.

to 115 per cent of its normal value. In the thyatron tube rectifiers, the circuit of Fig. 19 operates into a constant resistance load of several megohms. With such a high value of load resistance, the addition of the compounding resistor on the load side of the regulating resistor does not cause appreciable error. In fact, laboratory measurements on an experimental unit show that the compounding can be adjusted to obtain *improved* regulation of the thyatron tube rectifier when the grid battery eliminator is used in place of the normal grid battery. This is because the grid battery eliminator can be adjusted to over-correct for line voltage variations and thus compensate for the slight amount of residual line regulation error in the thyatron circuit.

The thermistor in Fig. 19 is a shunt element across one of the resistors in the potentiometer and a change of its resistance is equivalent to changing the potentiometer adjustment. The thermistor decreases its resistance with an increase of ambient temperature so it will change the output voltage when the temperature is changed. This output voltage change is opposed to the voltage changes resulting from the temperature effects in the reference voltage diode and the transistor. By selecting the proper thermistor and the proper ohmic values for the potentiometer resistors, these temperature variations will nearly cancel and the regulated output voltage will be temperature compensated.

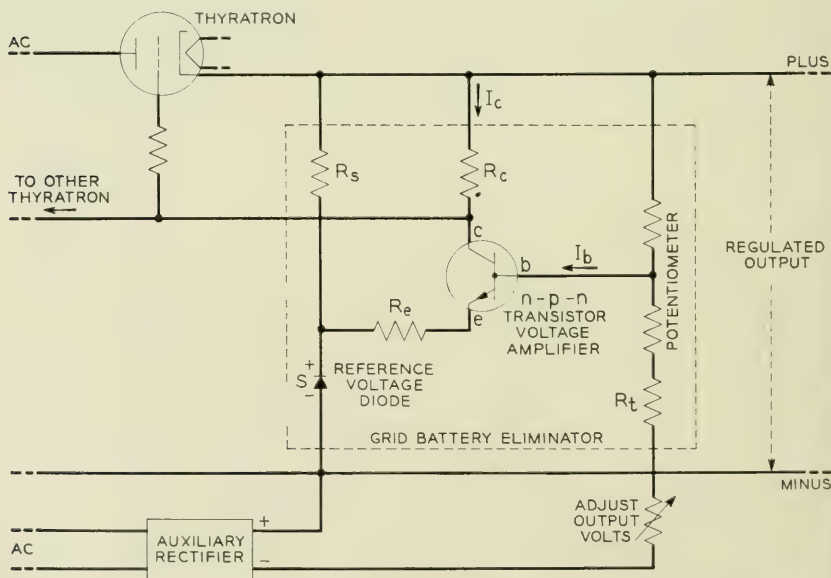


Fig. 20 — Grid-battery eliminator for magnitude controlled thyatron rectifier.

4.22. *Magnitude Controlled Thyatron Tube Rectifiers*

The grid battery eliminator covered in Section 4.21 is also usable in magnitude controlled thyatron tube regulated rectifiers but a simple, less expensive circuit can be used for this application. It is illustrated in Fig. 20. A simplified schematic of the thyatron rectifier is also shown in Fig. 20 and the grid battery eliminator is the portion of the circuit enclosed by the dotted line. It is actually a transistor voltage amplifier circuit. This type of circuit has not been covered previously in this discussion so its operation is described in some detail below.

Referring to Fig. 20 a portion of the output potential is compared to the reference potential by the base and emitter connections to the transistor. The difference between these two potentials causes the base current I_b to flow. This base current is amplified by the current gain of the transistor and it results in flow of collector current I_c , through the R_c collector resistance. The voltage drop across R_c is the negative grid potential applied to the thyatron tube. Now when the output potential is increased the base current is increased, the collector current is increased, the voltage drop across the R_c resistor is increased and the negative grid potential at the thyatron tube is increased. This will delay the firing of the thyatron and thus reduce the output potential.

If the ohmic value of the R_e resistor in Fig. 20 is zero the voltage amplification of this transistor circuit will be about 10, or a small change in the output potential will result in about 10 times this change in the thyatron grid potential. This is voltage amplification added to the circuit by the grid battery eliminator and a voltage gain of 10 is more than present circuits can use. The emitter resistance R_e reduces the voltage amplification of the grid battery eliminator to reasonable proportions.

The R_t resistance in the potentiometer circuit of Fig. 20 is wound with nickel resistance wire. Its positive temperature coefficient of resistance compensates the grid battery eliminator circuit for the temperature effects in the reference voltage diode and the transistor. This nickel wire resistance accomplishes the same result as the thermistor in Fig. 19. This is another method of compensating transistor regulator circuits for ambient temperature variations.

The "Adjust Output Volts" potentiometer and the auxiliary rectifier shown in Fig. 20 are part of the present magnitude controlled thyatron tube rectifiers. The auxiliary rectifier adds some ac line voltage compounding to the rectifier regulation. It is also used with a time delay relay circuit, not shown, to bias the grid potential of the thyatron tubes

similar to that in Fig. 18, except that an additional stage of current amplification has been added to the basic push-pull circuit.

Briefly, the regulating action is as follows. (1) The currents I_1 and I_2 respond in push-pull fashion to changes in output voltage V_I as covered in Section 3.36, (2) currents I_1 and I_2 are amplified by the 2-watt n-p-n transistors (T_3) and (T_4), (3) the amplified currents (I_3) and (I_4) flow in control windings (C_1) and (C_2) of the magnetic amplifier to control the voltage absorbed by the power winding (L_1), (4) this action regulates the average value of the voltage rectified by the germanium diode (D_1), thus completing the feedback loop. Tests show that this circuit is capable of ± 1 per cent accuracy of the output voltage with a ± 15 per cent change in the line voltage and with load current variations of from 10 to 100 per cent of rated output current.

4.32. 65-Volt, 200-Ampere Germanium Rectifier

Fig. 22 is a circuit sketch of a 65-volt, 200-ampere regulated rectifier suitable for charging and floating central office storage batteries. This rectifier employs six of the power rectifying cells with forced air cooling

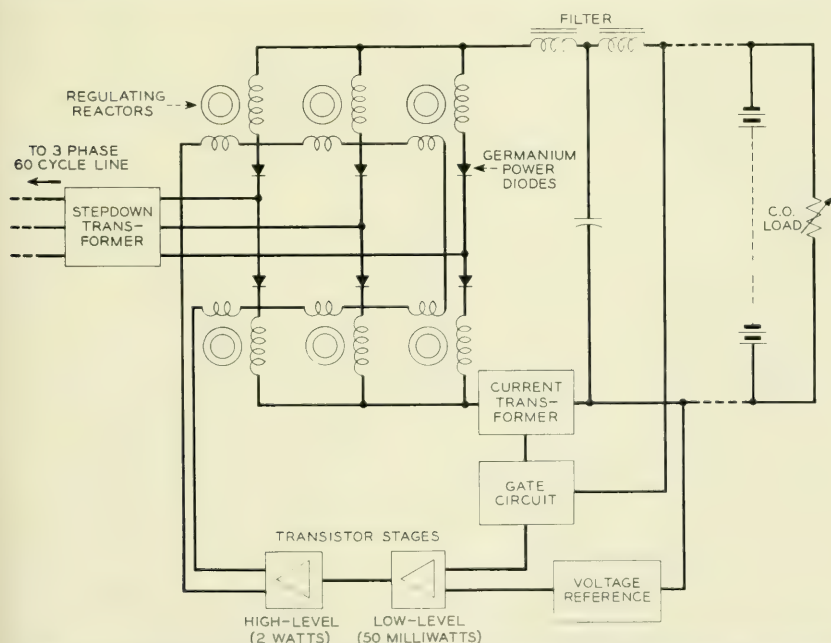


Fig. 22 Sixty-five volt two-hundred ampere magnetic amplifier regulated rectifier.

described earlier (Diode IV, Fig. 2), a reference voltage diode (Diode I, Fig. 2), two 50-milliwatt, and two 2-watt junction transistors.

The dc output voltage of the rectifier is controlled by a high gain self saturating magnetic amplifier. High gain in the magnetic amplifier is achieved by using tapewound gapless nickel-iron cores having rectangular hysteresis loops. The control current for the magnetic amplifier is provided by 2, 2-watt n-p-n transistors acting in push-pull. The 2-watt transistors are driven by 2, 50-milliwatt p-n-p transistors also acting in push-pull. The circuit is similar to Fig. 21. Again, the reference potential is furnished by a reference voltage diode.

Where the rectifier is connected to storage batteries an additional feature known as "current droop" is needed to protect the rectifier. The output characteristic of the rectifier with current droop is shown in Fig. 23. This characteristic is obtained by coupling a signal proportional to load into the first stage transistor amplifier through a gating circuit. This signal is provided by a dc current transformer which is another form of magnetic amplifier. At currents below the "droop" value the current signal is blocked from the amplifier. At full load the gating circuit allows the current signal to take over and hold the output current constant over a wide range of output voltage. In Fig. 23, the performance

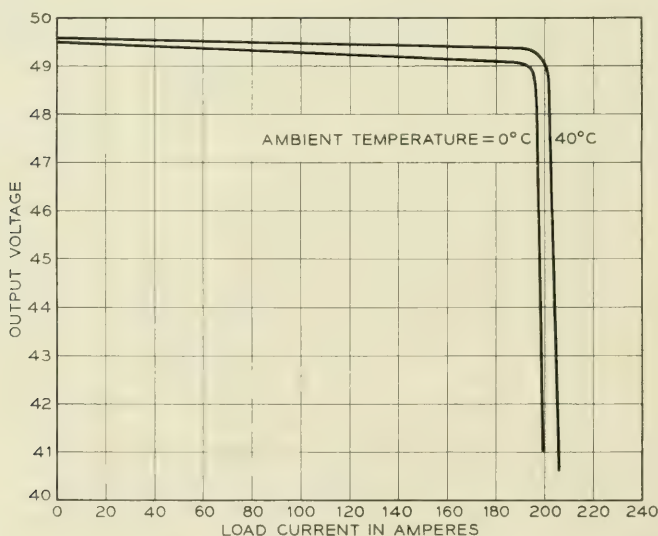


Fig. 23 — Output characteristics of experimental 65-volt 200-ampere germanium rectifier.

is shown over the range of ambient temperatures normally encountered in central offices.

5. CONCLUSIONS

It is seen from the above discussion that semiconductor junction diodes and transistors have a wide field of application in power conversion and control. Certain difficulties remain to be overcome, among which the variation of the device characteristics with ambient temperature appears to be the most troublesome at the present time. It has been shown that these variations with temperature can be minimized by two methods. First through the use of thermistors (negative temperature coefficient) or nickel-wire resistors (positive temperature coefficient) and second, through the employment of circuits in which the temperature variations of one element are balanced out by similar temperature variations in a complementary element. Thus, errors due to temperature changes can be minimized by further reduction of the sensitivity of the device characteristics to ambient temperature changes and by improved uniformity of the devices.

Another important aspect of the circuits covered in this paper is their freedom from dependence on auxiliary sources of dc potential. In most cases it is possible to power the regulating circuit directly from the regulated output, thereby eliminating the necessity for the transformers, rectifiers and filters usually needed to furnish plate potential for the regulating tubes and voltage standards.

The regulating circuits discussed in this paper are of the checkback type. In all of them, there must first be an error in the load voltage to start and maintain the regulating action. The load voltage will only return to precisely the original value if the regulating amplifier has infinite gain. These effects, however, are common to all closed-loop feedback regulating systems. Transistors and junction diodes, at their present stage of development seem well suited for use in checkback circuits having a high quality reference potential, for the feedback principle helps to minimize residual errors due to changes in the device characteristics with changes in ambient temperature.

Of course, the small size, long life and high efficiency of these semiconductor junction devices will also be very gratifying to the design engineers.

6. ACKNOWLEDGMENTS

The authors wish to acknowledge the assistance of C. W. Van Duyne of Bell Telephone Laboratories and the personnel in his group,

who were instrumental in obtaining the experimental data upon which this paper is based.

7. REFERENCES

1. W. Shockley, The Theory of p-n Junction in Semiconductors and p-n Junction Transistors, B.S.T.J., **28**, p. 435, 1949.
2. C. L. Rouault and G. N. Hall, A High-Voltage, Medium-Power Rectifier, I.R.E. Proc., **40**, p. 1519, Nov., 1952.
3. G. L. Pearson and B. Sawyer, Silicon p-n Junction Alloy Diodes, I.R.E. Proc., **40**, p. 1348, Nov., 1952.
4. R. M. Ryder and R. J. Kircher, Some Circuit Aspects of the Transistor, B.S.T.J., **28**, p. 367, 1949.
5. R. L. Wallace, Jr., and W. J. Pietenpol, Some Circuit Properties and Applications of n-p-n Transistors, B.S.T.J., **30**, p. 530, 1951.
6. R. N. Hall, Power Rectifiers and Transistors, Proceedings of the I.R.E. Proc., **40**, p. 1512, Nov., 1952.
7. W. Shockley, *Holes and Electrons in Semiconductors*, D. Van Nostrand, 1950.
8. K. G. McKay, Avalanche Breakdown in Silicon, Phys. Rev., **94**, p. 877, May 15, 1954.

Wire Straightening and Molding for Wire Spring Relays

By A. J. BRUNNER, H. E. COSSON and R. W. STRICKLAND

(Manuscript received January 19, 1954)

The basic design of the wire spring relay departs from conventional relay design in many ways. Translation of some of these design departures into commercial relay manufacture has necessitated the development of new machines and new methods because those available were incapable of producing to the new design requirements. Two developments in this category involved the straightening of large quantities of small diameter wire and the molding of a multiplicity of straightened wire inserts into phenolic resin blocks. The manner in which these developments were reduced from problems to practice is the subject of this paper.

PART I — AUTOMATIC WIRE STRAIGHTENING

Ordinarily wire is received from suppliers on spools or reels. In the spooling operation a spiral bend is placed in the wire which persists when it is unspooled. For use as a wire spring in the wire spring relay this spooling bend must be removed if the wire is to be positioned with the precision required for the desired functioning of the relay. It is necessary, also, to have the wire free of bends if automatic manufacturing methods are to be employed. For these reasons, it is important that the nickel silver and silicon copper wire used in the wire spring relay be straightened as the initial operation in the manufacture of wire block assemblies or "combs" for these relays.

Wire straightening can be accomplished by cold working the wire under controlled conditions until sufficient stress has been built up, particularly at the surface, to make the wire resist bending efforts. The degree of straightness required is governed, of course, by the desired performance of the comb in the operation of the wire spring relay. For the 0.0226-inch nickel-silver wire used in the twin wire comb this has been established, for example, as a deviation not exceeding 0.010-inch from absolute straightness measured at the contact end of the comb,

i.e. $2\frac{5}{8}$ -inches from its anchorage point in the phenol resin block. This degree of straightness is satisfactory also for the automatic manufacture of relay combs in which a multiplicity of straightened wires are guided into a molding die and positioned so accurately that they can be permanently imbedded in phenolic resin to the close dimensional limits necessary for ultimate assembly into relays.

EXPERIMENTAL WORK

Wire Straightening

The original experimental work on wire straightening was done at the Bell Telephone Laboratories to aid in establishing the feasibility of a wire spring relay design. After eliminating other approaches it was decided to straighten the wire in a motor-driven machine by pushing the wire through carefully oriented dies in a rotating head. The wire produced in this manner was known to have a twist but was adequate for making model parts. Subsequently, Western Electric development engineers made a survey of available commercial wire straightening machines. A machine was purchased which, while not intended for straightening wire of the small diameters used in wire spring relays, was capable of modification. Among the important things learned from the operation of this machine were first, it is preferable to push instead of pull wire through the rotating die head because of interference at the puller due to twist in the straightened wire; second, much of the twist can be removed from the straightened wire by spinning the spool of raw wire counter to the direction of the driven die head; and third, it appeared that a simpler approach than spinning the spool of raw wire would be to pass the wire through a second straightening head rotated in the opposite direction from the first. On the basis of these observations, a Hawthorne-designed experimental straightening machine was constructed. This machine featured two die heads independent of each other and counter rotating in operation. Five individual sets of die blocks, with provision for spacing adjustment as found on the rotary die holder of the commercial straightener, were retained in each head. Subsequently, this experimental machine was used for an extended series of tests to determine such things as the optimum spacing between individual dies, the proper offset from the center line of the head for each die, the best ratio of opposing head speeds, and the maximum rate of wire feed with respect to the rotational speed of the die holder head required to produce straight wire in the diameters employed in the wire spring relay.

Since it had become evident by the time this study was well advanced

that a multiple head machine would be required, a second experimental machine was built. This machine, Fig. 1, designed with the driving mechanism and spacing allowances considered necessary for an automatic multiple head straightener, had only one double head capable of straightening a single wire. A major change, to be discussed later, was replacement of the five adjustable die blocks in each head by a pair of opposing die blades contoured to provide a wire passage space between them identical to the predetermined path previously forced upon the wire by the five die sets. These die blades were retained by a spindle keyed to the drive mechanism. To accommodate the double head feature, a rotating unit consisting of two spindles coupled together was employed. Much of the remaining experimental work, such as optimum rotational speed of the spindles, the effect of different configurations of the die blade wire path surfaces, rate of wire feed, etc. was performed with this machine. Except for minor changes it became the prototype for the automatic multiple head machines constructed later.

Straightened Wire Storage

In contrast to the manual operations needed to assemble the springs and phenol fibre insulators of U- and Y-type relay spring pile-ups, it was planned from the beginning to mold straightened wire into phenolic resin by automatic means so that unit assemblies would be obtained for

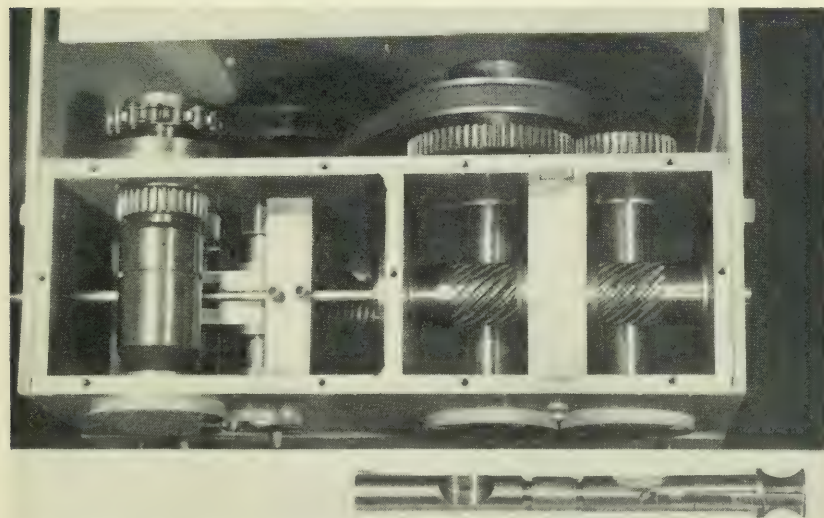


Fig. 1 — Experimental machine with one double head, prototype of 24 and 30 double head wire straightening machines.

the wire spring relay. The labor economy of the latter procedure is obvious. To implement this plan it was necessary that straightened wire be available at the molding press in the quantities required to prevent loss of molding time. To be successful it was important that interruptions to the regular recurrence of molding cycles, such as rethreading the multiplicity of wires into molding dies, be kept to a minimum.

The original planning envisioned a battery of single strand wire straighteners operating continuously to make relatively long lengths of straightened wire. How to store this wire between the wire straightener and the molding press presented the real problem. An early attempt toward a solution involved winding straightened wire on 36-inch diameter reels until sufficient length had been accumulated for eight hours' molding time. These reels would be mounted ahead of the molding press as shown in Fig. 2, and changed at the end of each eight hour shift.

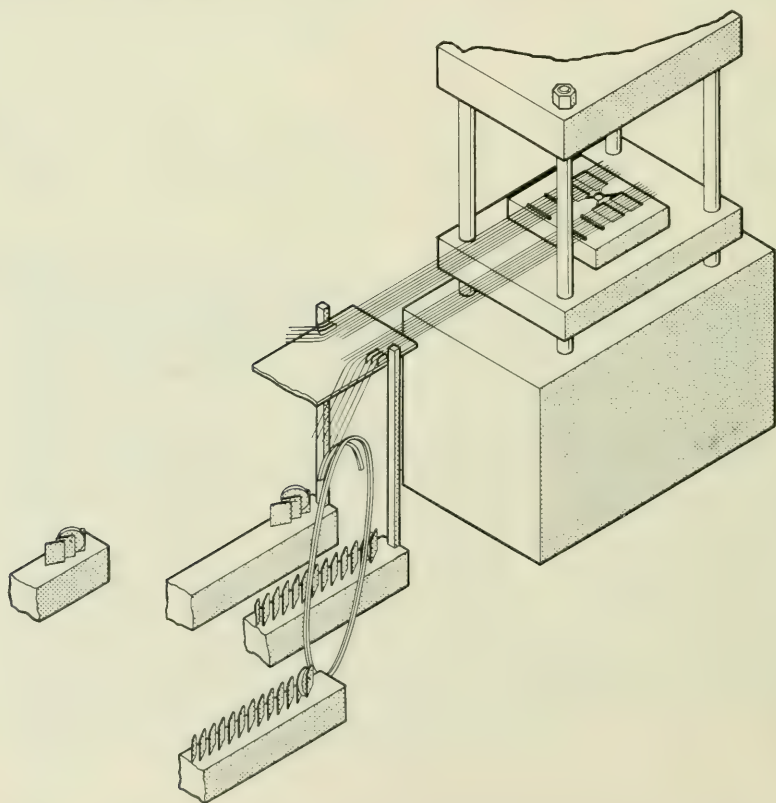


Fig. 2 — Sketch showing handling of straightened wire on storage reels.

Initial efforts indicated that this procedure was practicable. However, a new shipment of nickel silver wire revealed that, while not detectably different from previous shipments, the new wire took a permanent set on the 36-inch reels thereby making it unusable at the molding press. Principally because of the incipient possibility of straightened wire acquiring a set when not stored on flat surfaces, this reel approach was abandoned. Another effort consisted of providing a multiplicity of straight storage tubes, of either metallic or plastic material, into one end of each of which an eight hour supply of a single strand of wire was pushed by the wire straightening machine and from the other end of which a molding press would withdraw its requirement of wire (Fig. 3). This was found unworkable because often the wire straighteners were unable to push the required length of wire into the tubes due to the lead end becoming snarled from twist in the wire. It was decided, finally, to discard the idea of continuously straightening and storing wire in favor of placing multiple head machines adjacent to the molding presses and operating them only as required. This meant increasing the straightening machine investment because intermittent operation of the straighteners necessitated more wire straightening facilities. A compensating factor was the elimination of investment in storage facilities. It was found that interrupting the continuous operation of the straightening heads had no detectable effect on the characteristics of the straightened wire. Accordingly, multiple head automatic wire straighteners are now placed adjacent to the molding press and operated at a speed slightly greater than the wire consumption of the molding dies. Automatic control of the length of a partial loop of wire extending from the wire straightener assures an adequate supply of wire at the molding press. The ultimate length of continuous straightened wire available to the molding press by this arrangement is governed only by the length of raw wire on the spool and is sufficient for many operating shifts.

AUTOMATIC MULTIPLE HEAD WIRE STRAIGHTENER

Both 24- and 30-double head automatic wire straighteners have been built by the Western Electric Company. The 24-double head straighteners are used in making combs for the AF, AG and AJ type general purpose wire spring relays and the 30-double head machines for the 286, 287 and 288 type multi-contact relays. In practice the phenol resin molding operation is accomplished in four-cavity dies with the cavities arranged symmetrically about the center of the die. Thus two forward cavities face the wire straightener with the remaining two cavities in tandem. When making the twelve wire single comb of general purpose

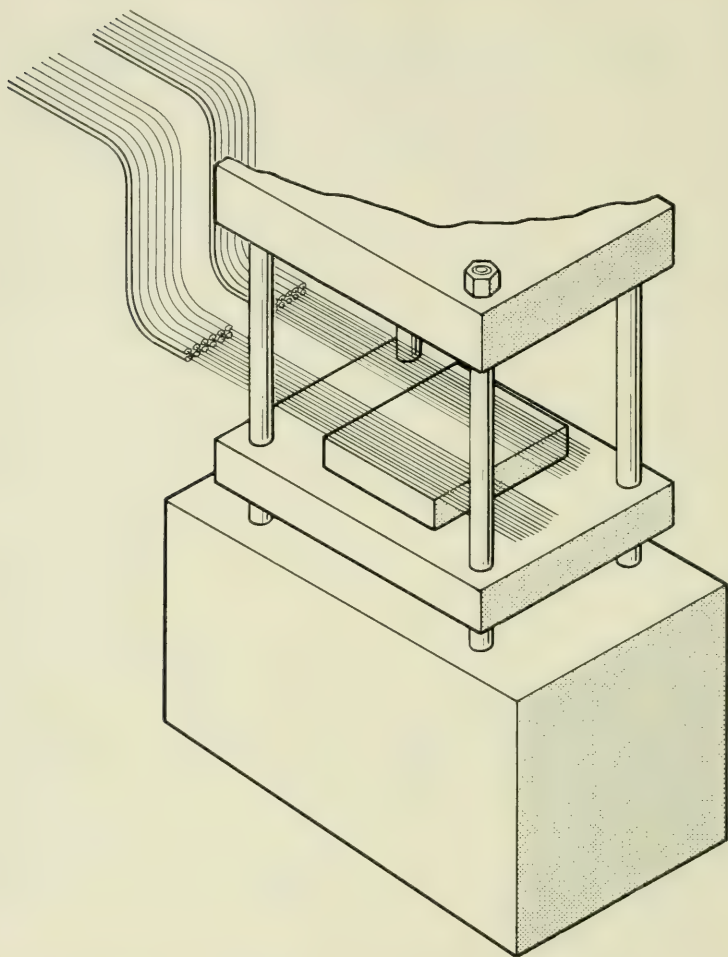


Fig. 3 — Sketch showing method of handling straightened wire from storage tubes.

relays, half of the wires from the 24-double head straightener are guided into each forward cavity. When the twenty-four wire twin wire comb is being made, on the other hand, the entire production of a 24-double head straightener is guided into each forward cavity and two wire straighteners are necessary for each molding press.

The 30-double head straighteners are arranged similarly when the fifteen wire single wire comb and the thirty wire twin wire comb of the multi-contact relay are being molded.

WIRE SUPPLY

One spool of raw wire is cradled in the wire straightener, Fig. 4, for each wire required in the molded comb. There are three sizes of wire straightened, 0.0200 and 0.0226-inch diameter wire for the twin wire comb of multi-contact and general purpose relays, respectively, and 0.0400-inch diameter wire for the single wire comb of both relays. The smaller wires are nickel silver while the 0.0400-inch wire is a silicon-copper alloy. All three wires are in the hard temper range.

Originally the wire was pulled from the spools by the drive (pusher) roll of the wire straightener. However, the pulling force required varied widely from spool to spool. The result was an unequal rate of wire feed through the straightening heads. To avoid this, a capstan with an individual pulley adjustment for each wire was added to the machine. This capstan, in addition to pulling the wire from the supply spools, meters the amount of wire fed into the straightener. An occasional adjustment of individual capstan pulleys is all that is necessary now to assure production of straightened wire at a uniform rate from every head.

WIRE STRAIGHTENING MECHANISM

Fig. 5 shows the wire straightening mechanism. Some of the spools of raw wire are visible to the right below the 24 wires, in this instance, being pulled from the capstan pulleys by the grooved shaft mounted just inside the machine housing. This shaft has 24 grooves, one for each wire, which mate with twelve spring tensioned twin grooved wobble rolls underneath to provide the means for pushing the wires through the straightener heads. Both the grooved shaft and the twelve mating rolls are power driven. The wires are pushed through the tubes to the left of the grooved shaft which guide them to the spindles in the straightening heads. These heads, arranged in two vertical rows, make it possible for a common spiral geared drive shaft to rotate all 24 heads at identical speed. This arrangement, however, causes twelve of the heads to rotate clockwise and twelve to rotate counterclockwise. The opposite twists produced in the upper and lower wires under this circumstance are corrected for by the double head arrangement in which the second set of heads rotate counter to the first set.

DIE BLADES AND SPINDLES

Inside each head is a removable spindle for retaining a pair of contoured wire straightening die blades. The spindles are suitably keyed

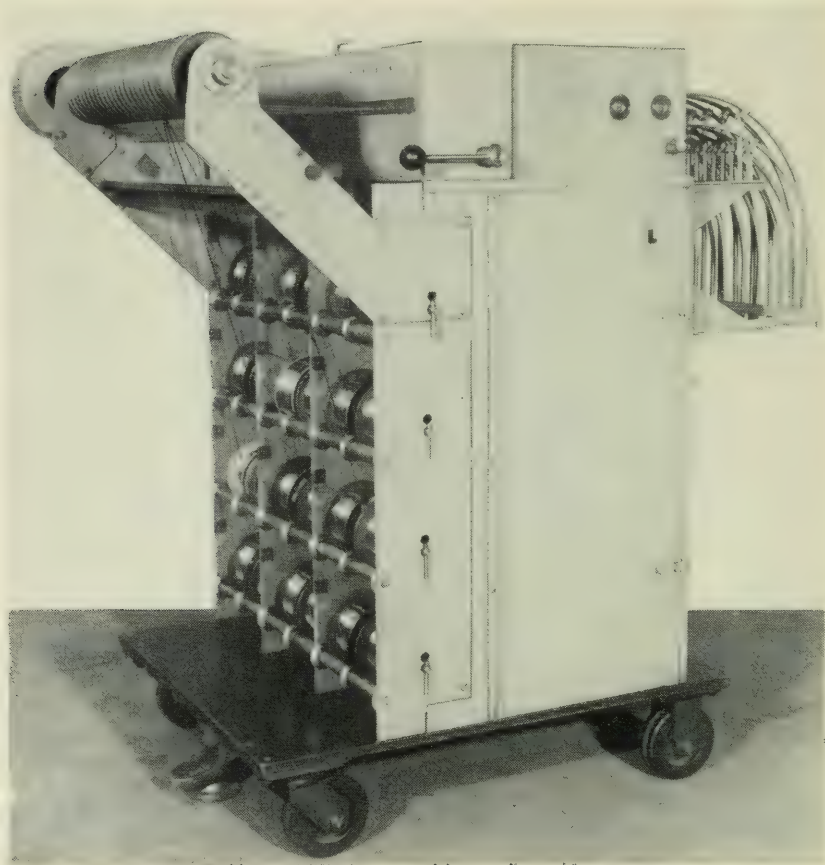


Fig. 4 — 24 double head wire straightening machine.

to the heads to assure rotation. To conform to the double head design two spindles, joined by a loose coupling, are used. This is illustrated in Fig. 6 which also pictures two sets of die blades removed from the spindle slot. The space between each pair of contoured die blades as positioned for the photograph shows clearly the path of the wire during its transit of the rotating heads.

The die blades maintain the same spacing, offset and length that constituted the desired wire path through the five individually adjusted sets of die blocks of the early straightening machines. The continuity of a die blade is accomplished simply by bridging what had been air spaces between the individual die elements and removing enough metal to prevent wire contact in the bridging sections.

The die blades are used not only to conserve space but also to minimize die costs. The latter is accomplished by making them from inexpensive sheet metal on a punch press and discarding them as soon as wear has destroyed their usefulness for wire straightening. Unlike the individual die blocks used in the previous rotary head straighteners, it is not necessary to groove these die blades to direct the flow of wire through the head. There is a slot milled into each spindle to hold the die blades as shown in Fig. 6. The walls of these slots guide the wire and limit its sideways movement in much the same manner as the grooves in the individual die blocks. The actual thickness of the die blade was established as slightly more than that of the diameter of the largest wire to be straightened for wire spring relay combs. Thus, one slot of uniform width is milled into each spindle allowing interchangeability of spindles regardless of the diameter of the wire to be straightened.

Wire in its transit through the die blades is flexed and burnished to the extent required to produce the desired degree of straightness. It is not rotated during the straightening operation but may acquire twist and even a spiral threadlike burnished appearance from rotation of the die blade surfaces.

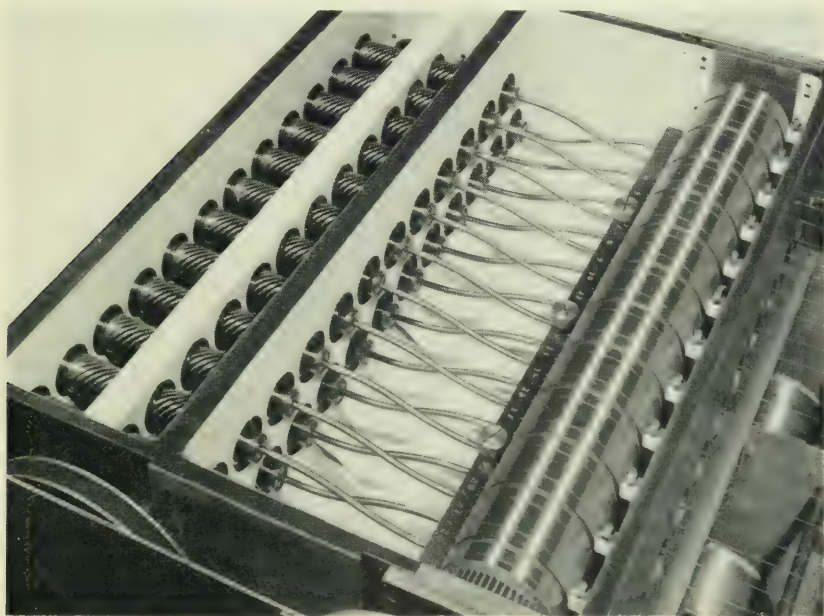


Fig. 5 — Straightening mechanism of 24 double head wire straightening machine.

OBSERVATIONS ON WIRE STRAIGHTENING

The straightening operation affects some physical properties of the wire. Tensile strength is reduced about 10 per cent while elongation is increased around 50 per cent. The diameter of the straightened wire is usually from 0.5 to 1.0 per cent greater than that of the raw wire with commensurate loss in wire length. Both straightness and twist appear to be dependent in large part upon the contours of the die blades. Thus far these contours have been determined by trial and error on the adjustable die block straightener, although general relationships, especially with respect to wire size, are becoming evident. It is expected that further study and experience will establish bases on which contours can be calculated with accuracy.

Twist imparted to the wire by the rotating action of the spindle has been found difficult to measure. What is referred to as twist is actually

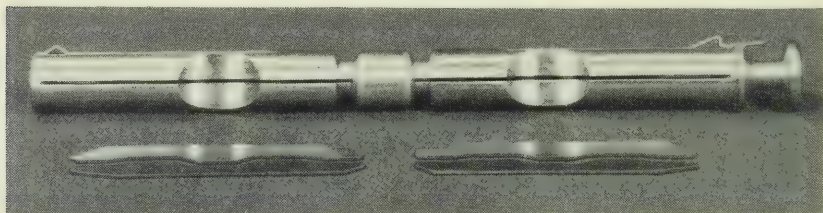


Fig. 6 — Double spindle showing slots and complement of two sets of die blades.

radial distortion of the wire about its longitudinal axis resulting from partial release of internal stresses remaining in the wire after straightening. Further release of internal stresses may occur when the wire ends of the twin wire comb are formed before welding, in which event mislocation of contacts will result, Fig. 7. This is objectionable from the standpoint both of subsequent manufacturing operations and of relay performance. The internal stresses are caused by the crank action applied to the wire surface while it is passing between the die blades in the rotating spindles. Internal stress which is not apparent until after its release, as by forming, has been designated as "residual twist".

A rough approximation of the amount of residual twist in wire can be obtained by measuring what has been termed "apparent twist". Apparent twist is the amount of visible rotation at the end of a wire after leaving the straightener. It can be measured in degrees of rotation per foot of wire straightened. When the apparent twist is low, usually the residual twist also is low. A working range for permissible apparent

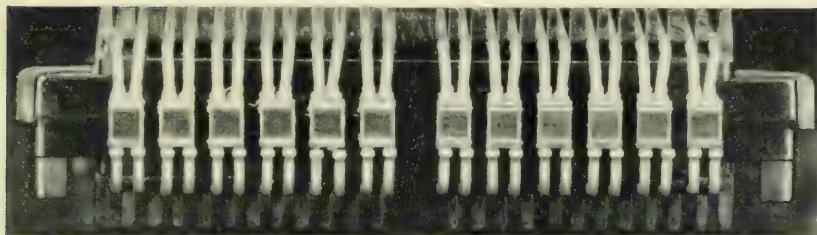


Fig. 7 — Photograph showing the affect of residual twist in the upper set of wires as compared to freedom from residual twist in the lower wires.

twist has been established which has been successful generally in maintaining acceptable limits on residual twist.

There are three variables which largely control the quality of straightened wire. The first is the physical properties of the wire itself. Although a shipment of wire may, on the basis of the sampling method employed, meet specification requirements limiting physical and chemical characteristics, an occasional spool or part spool of wire can be expected which will be enough outside limits to cause unsatisfactory straightness and unmanageable residual twist.

The second variable affecting straightness and twist in wire processed on multiple head machines lies in small differences between the rotating head assemblies. While all critical dimensions of the die blades, spindles, and spindle housings are held to close tolerances, it is possible to obtain an accumulation of dimensional deviations in some spindle assemblies of such magnitude as to cause appreciable difference in wire twist and sometimes in wire straightness. It has been necessary, therefore, to provide means for balancing such dimensional variations.

The third variable is wear on working surfaces of the die blades. Continued sliding of wire over die blade surfaces eventually produces grooves which decrease offset and increase clearance in the wire path. It has been found in general that, as the die blades wear, twist decreases until it eventually reverses direction. Simultaneously, straightness may improve to a critical point from which it rapidly deteriorates. Accordingly, replacement of die blades must be made before wear has rendered them ineffective.

CONCLUSIONS

Satisfactory performance of the multi-head wire straightening machines described has been demonstrated during the pilot plant period of wire spring relay manufacture. Further refinements in the means

for controlling known variables must be made, however, to assure the reliability demanded of heavy duty mass production machines.

PART II — AUTOMATIC MOLDING OF WIRE SPRING RELAY BLOCK ASSEMBLIES

Parallel with the effort directed toward development of wire straightening facilities, an investigation was undertaken by Western to develop automatic facilities for molding an array of straightened wires into small plastic blocks spaced at specified intervals. These blocks were designed not only to hold the wires securely and to locate them accurately but also to insulate them from each other electrically. The design engineers at Bell Telephone Laboratories had decided, after evaluation of the physical properties of available plastic molding materials, that a thermosetting phenolic type resin would best provide the characteristics needed for wire spring relay block assemblies. Proceeding on this information, Western Electric development engineers reviewed the merits of molding methods adaptable to embedding a multiplicity of inserts, wires in this instance, into phenolic resin. Such economic factors as molding time and material cost were balanced against molding problems like shrinkage and flow characteristics. It appeared from this review that transfer molding offered the most favorable possibilities. It appeared also that the shortest practicable molding cycle might be achieved by preforming the phenolic resin material, preheating these preforms electronically and automatically feeding them into the molding die. Further study of the molding problems indicated that molding presses for this purpose would have to be specially designed, particularly if multicavity dies were to be used. The special design features, such as wider spacing between the tie rods, provision of an electronic preform heater, and micro-timing devices, are discussed later as they become pertinent to the description of the machines finally adopted.

AUTOMATIC MOLDING MACHINE

Hydraulic molding presses appeared to offer the most advantages for this job. Essentially, these consist of two opposed hydraulic rams mounted vertically; the lower and more powerful ram providing the force required to close and hold closed the split die employed and the upper ram providing the force needed to transfer the phenolic resin in a softened or plastic state into the die cavities.

Unusually wide spacing between the tie rods of the press had to be provided to accommodate the complex progressive die required to make

the molding operation automatic. This die had to be designed not only to mold resin in multiple cavities but also to remove the plastic cull, index the molded blocks, and shear the completed assemblies. Provision had to be made, in addition, for mounting an electronic preform heater and for space to install an appropriate conveyor from the heater to the plastic transfer point of the die. These features were incorporated into an experimental prototype machine, shown in Fig. 8, operated under laboratory conditions. Adjustable microtiming devices were engineered into this machine to control the sequence of operations precisely and automatically.

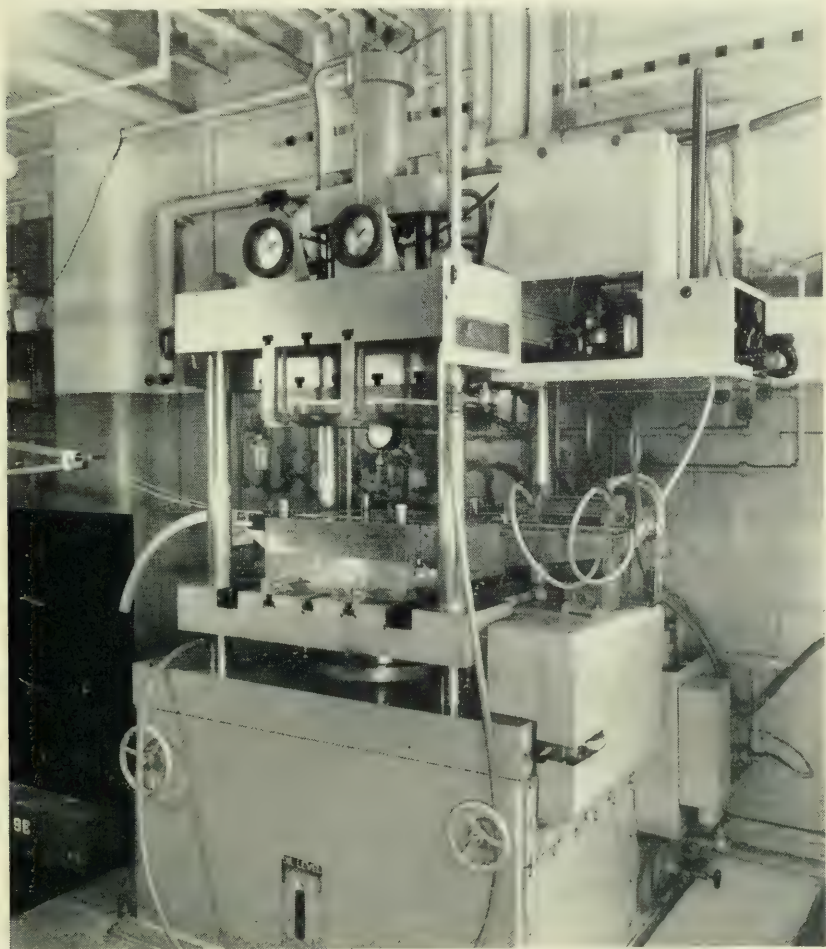


Fig. 8 - Experimental prototype automatic transfer molding machine.

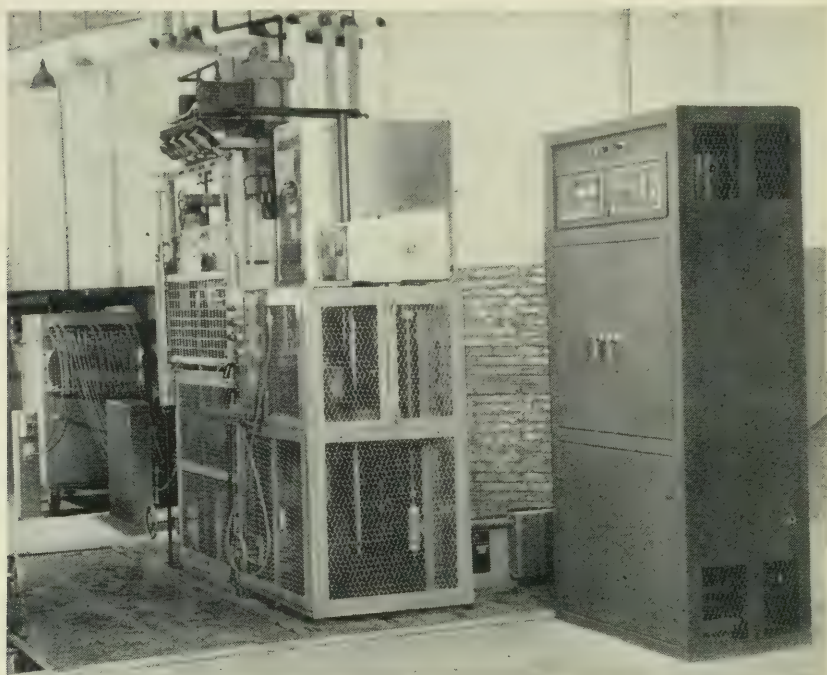


Fig. 9 — Typical installation of wire straightening and molding machines.

The final embodiment of the molding machine is shown in Fig. 9. The wire straightener included in the photograph is positioned a short distance from the molding machine to allow straightened wire to form a partial loop between the two. This permits the wires to leave the straightener at a constant speed from a fixed position and to enter the die at intervals controlled by the molding cycle and move vertically with the opening and closing motions of the lower half of the die. The straightening machine is started and stopped by an electrical control which assures the desired size partial wire loop at all times.

PHENOLIC RESIN PREFORMS

To operate on an automatic basis, it is important to have the molding compound in such form that it can be handled easily and that the charges are of uniform weight and size. This is done by compressing the bulky granular compound as received from suppliers into small carefully dimensioned cylinders or "preforms". Mechanical presses, capable of turning out preforms in multiple at each stroke, are used at Hawthorne

for this purpose. To obtain uniformly low water content in these preforms, it is necessary to store them in an air conditioned chamber for a three week period to assure attaining equilibrium conditions.

PREFORM HEATING

An electronic preform heater is a means of increasing molding machine production by shortening the time the phenolic resin must be retained in the molding die during each cycle. This is done by adding to the preform, just before it enters the die, much of the heat required to plasticize the resin. Thus, as one charge of compound is being molded in the die cavity, another is being preheated as part of the molding cycle. The amount of preheat that can be permitted is limited by the extent to which heat induced chemical reaction can be tolerated outside the molding cavity and varies with the size and contour of the die cavity. The rate at which the preform is heated influences its consistency at a given temperature.

A feed mechanism has been provided to convey the preform from a magazine to the electronic heater and thence to the die. This mechanism consists of a horizontal guide plate extending between the two grids in the upper part of the press. The preform is pushed along this guide plate by a metal bar mounted on endless roller chains at each side of the guide plate. In operation, a preform from the magazine is pushed to a position between the electrodes of the dielectric heater. The push bar is then backed away a small distance so that it will not interfere in the inductive field. Upon completion of the preheating operation, the push bar shoves the heated phenolic preform to and beyond a drop off position above the open end of the transfer cylinder of the molding die. The conveyor continues to operate until the push bar has removed a new preform from the magazine and loaded the heater in preparation for the next cycle.

PRESS CONTROLS

The electrical controls or "brain" of the production unit are housed in a cabinet adjacent to the molding machine. The operation of the press, the electronic heater, the feed mechanism and the pneumatic device on the die are all coordinated into precise sequences by micro-adjustment of these controls, Fig. 10. Any operational sequence or length of cycle desired can be established for repetitive manufacture. On the other hand, the press and the electronic heater can be placed on manual control at any time.

THE MOLDING DIE

Experimental Work

No attempt will be made to discuss the many ideas on mold design which were conceived, evaluated and either discarded or improved in arriving at the designs now employed. The investigation was undertaken

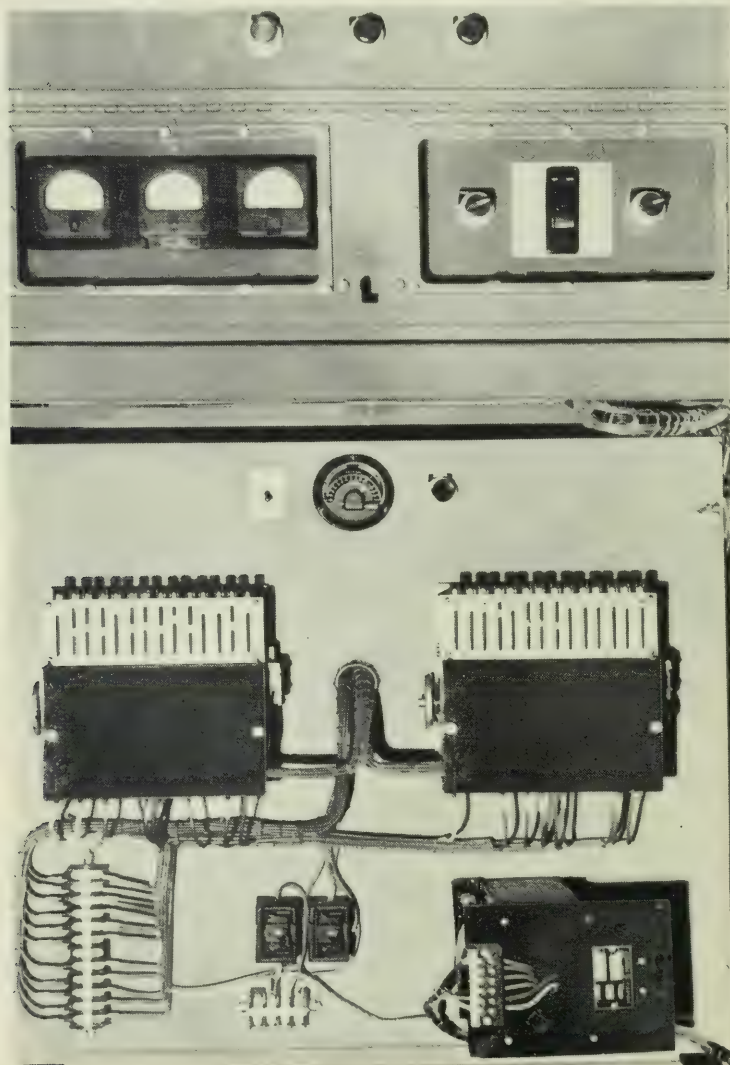


Fig. 10 Cycle timing and electronic heater controls for molding machine.

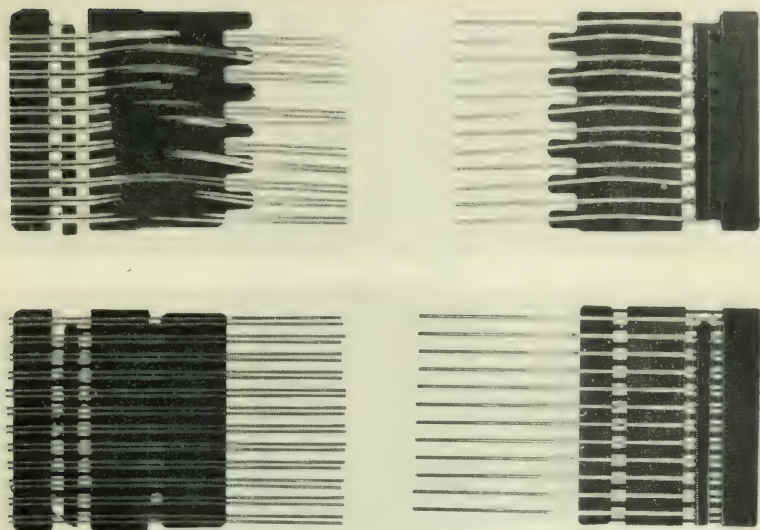


Fig. 11 — Cutaway sections of resin blocks showing inadequate versus adequate wire supports in molding die.

with no significant experience in molding closely spaced arrays of small diameter wires into phenolic resin. The initial effort demonstrated convincingly, Fig. 11, that small diameter wires cannot be embedded in resin by high pressure molding techniques without the liberal use of wire supports. These are needed to prevent individual wires from being deformed by the pressure of the plastic as it is forced into the cavity. One fundamental observed in die design subsequently was to keep all wire spans inside die cavities as short as possible.

As expected, early studies demonstrated that lack of cross sectional symmetry required combs to have a marked tendency to warp. While warping could be reduced by increasing the time the resin was retained in the molding cavity, this partial solution was unsatisfactory from the standpoints both of warpage and product cost. Accordingly, every effort was made, consistent with relay design requirements, to depart as little as possible from symmetrical die cavity design and where symmetry could not be achieved, to attempt to distribute the resin mass uniformly on each side of the center line of the wires.

The importance of symmetry, together with the desire to keep molding flash to a minimum, influenced Western development engineers to design the earlier experimental die cavities with the wire inserts centered at the

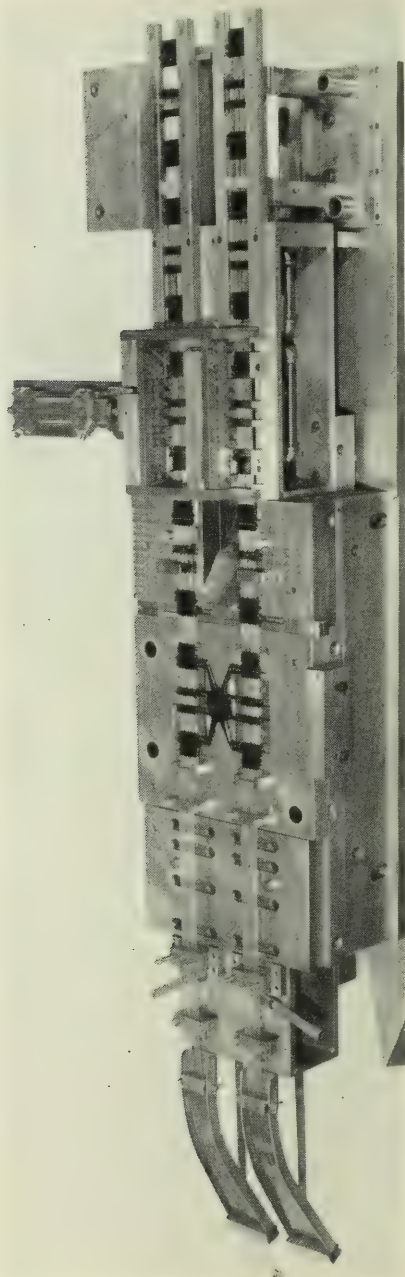


Fig. 12 — Lower half of typical automatically operated four cavity die.

die parting line. In the course of die development work it became apparent that reduction in both die cost and die construction time could be effected in some instances by confining the grooves for locating and supporting the wire array to one half of the die cavity. This type of design had the added advantage of eliminating annoying problems relating to precise registration of the upper and lower die halves during the molding operation.

The design of the die cavities now used for high production molding represents development work based on such considerations as those outlined above. The die cavities are similar but not the same for both twin and single wire combs. They differ in the location of the parting line which is flush with the top of the wire array in the twin wire comb die and principally at the center line in the single wire comb. The latter arrangement was dictated by two considerations, (1) Use of a U shaped groove in one half the die cavity for a wire as large as 0.040-inch diameter, resulted in excessive molding flash and (2) The chance of obtaining completely filled fins on both sides of the forward block was enhanced by centering the wires.

THE DIE

Fig. 12 shows the lower half of the complex automatically operated four cavity die which was evolved from simple single cavity hand loaded units. That this evolution may not be completed is indicated by the fact that the section of the die operated by the air cylinder shown at the top of the photograph and intended to remove molding flash from the wires is no longer used. Much better flash removal is effected by blasting the combs with ground walnut shells after the molding operation has been completed. The operation of the automatic die, as related to the manufacture of single wire combs containing twelve 0.040-inch silicon copper wires, will be described under five sub-heads in the same sequence as the progression of the wire array through the five sections of the die. Similar progressive operations are performed in making twin wire combs. The die consists of an upper half attached immovably to the head of the press and a lower half mounted on a hydraulic ram which raises and lowers to close and open the die.

WIRE GUIDING

The single wire comb die uses one 24-double head wire straightener. Fig. 13 shows how twelve continuous strands of wire from the straightener are guided to the required spacing in each comb array. Spring ten-

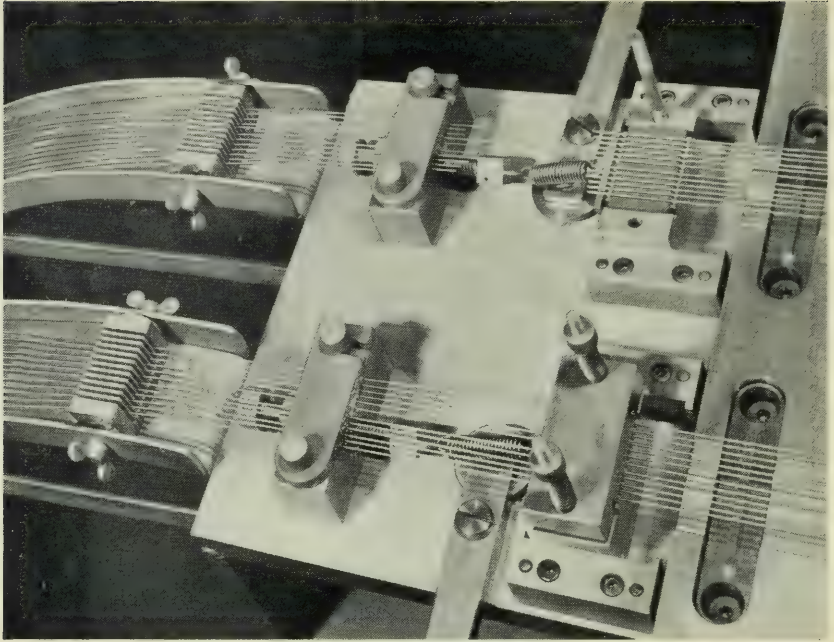


Fig. 13 — Wire guide section of automatic die.

sioned sliding blocks with spring loaded felt cleaning pads hold the wires taut during the operations that follow.

ANCHORING

Because phenolic resin does not wet and, therefore, does not bond with most metals, mechanical means must be used to prevent the wires from turning in the resin block of the finished combs. This precaution is necessary because wire wrapping tools are employed in wiring relays into equipment. The wrapping tool puts a torque on the wire which may, if the wire is not securely anchored, turn it in the resin block. Any movement, obviously, will mislocate the contact welded onto the wire and cause maladjustment of the relay. To mechanically prevent turning, a section of each half of the die is provided with accurately spaced mating blade inserts, Fig. 14, which press flat lands or "anchors" on the wire each time the die is closed. When the wires are indexed subsequently these anchors are located on that portion of the wire embedded in the resin.

MOLDING

Dies with four cavities are used for all comb molding operations. The cavities are located around the plastic transfer area as illustrated in Fig. 15. Above the transfer area a cylindrical opening extends vertically through the upper die half to permit passage of the transfer ram. By closing the die, the transfer area is enclosed except for small orifices leading to each of the four cavities. The heated preform is dropped into the center of the transfer area ahead of the transfer ram. This ram, operated by hydraulic pressure and heated by contact with the hot die, forces or transfers much of the heat softened resin through the orifices and into the die cavities. The resin residues which are left in the passages between the ram and the cavity gates upon completion of the molding cycle are called runners. The heat of the die, approximately 360°F ., further softens the plastic resin enabling the pressure of the ram to force it into intimate contact with all die cavity surfaces. Continued application of heat and pressure hardens or cures the resin by the process of polymerization, after which it is permanently infusible.

To successfully operate on an automatic basis, the cluster of wire, plastic blocks, ram slug, and runners must be removed from the molding cavity. This is complicated in the single wire comb because the plastic block at the contact end of the wires has very thin fin sections

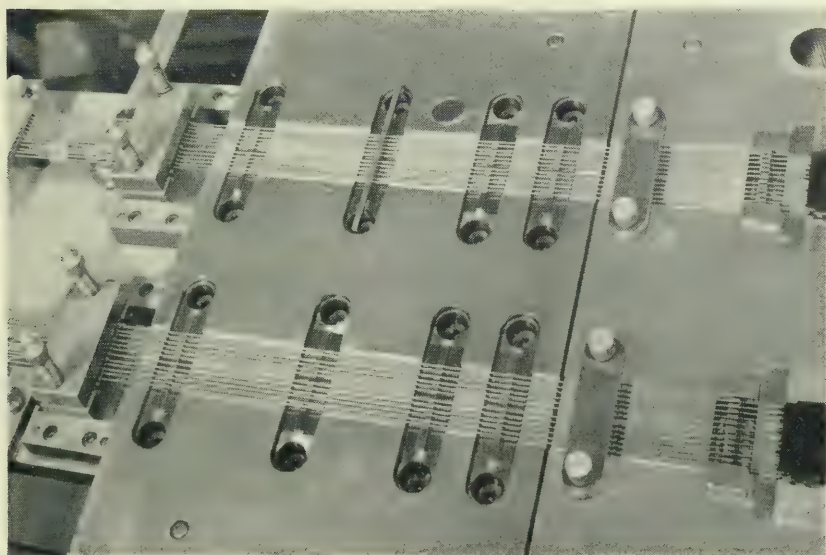


Fig. 14 Anchoring section of automatic die.

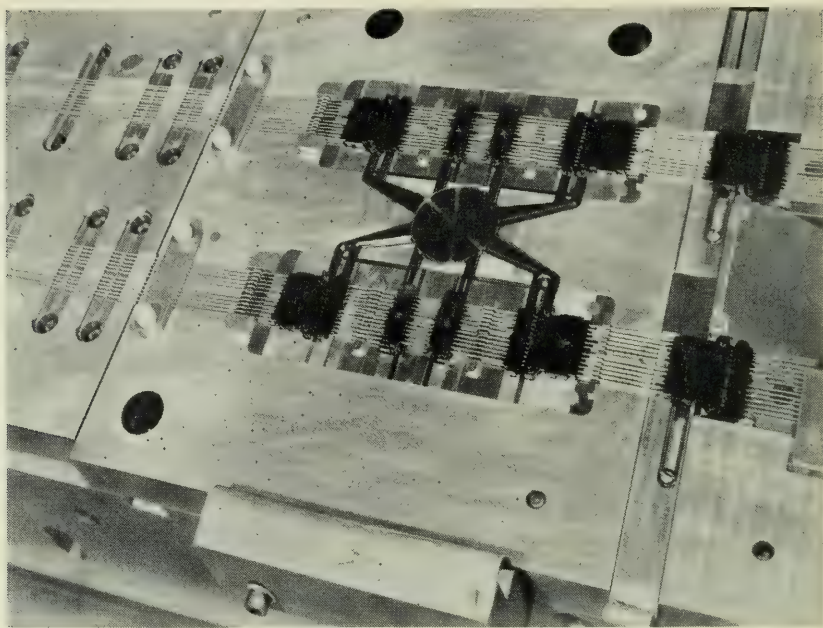


Fig. 15 — Molding section of automatic die.

projecting into both halves of the die cavity. These fins, used to guide the mating twin contact wires in the finished relay, must be held to close limits dimensionally and are thin to the point of fragility. To insure satisfactory removal of the resin blocks from the upper half of the die while simultaneously retaining them in place in the lower half when the die is opened, spring loaded ejector pins in the upper die half push down against the cured blocks. The ejector pins operate until the wire-resin cluster has been ejected from the upper die. Subsequently they are retracted by contact with reset pins which butt against the lower die half when the die is closed. The openings in the die half which accommodate these ejector pins serve as air vents when the resin is entering the cavity. To prevent the plastic cull, i.e., the resin slug and associated runners, from breaking off and damaging the die on the next molding cycle, the pressure of the transfer ram is maintained on the slug until the upper die surface has been cleared.

The hydraulic ram bearing the lower die half continues to withdraw until the lower ejector pins can function. These ejector pins are more numerous and more complex in design than those in the upper die half because, in addition to ejecting the resin blocks from the die cavities,

they aid in guiding the progression of wire inserts through the die and in locating the incoming wires in the die cavities. The operation of this ejection or "knock-out" consists of freeing the wires and resin blocks from the die cavities and then moving the newly molded cluster horizontally from the lower die cavity. When the die surface has been cleared by this indexing operation, compressed air is blasted across the die to remove loose molding flash which might be present. With the completion of the indexing operation, the ejector pins are restored to their original position by spring pressure. As a precaution, reset pins are provided in case this spring pressure should be inadequate.

COOLING AND CULL REMOVAL

The next operations are those of cooling the plastic to minimize warpage and removing the cull. Because the larger plastic block of the single wire comb has surface irregularities on one side and is smooth on the

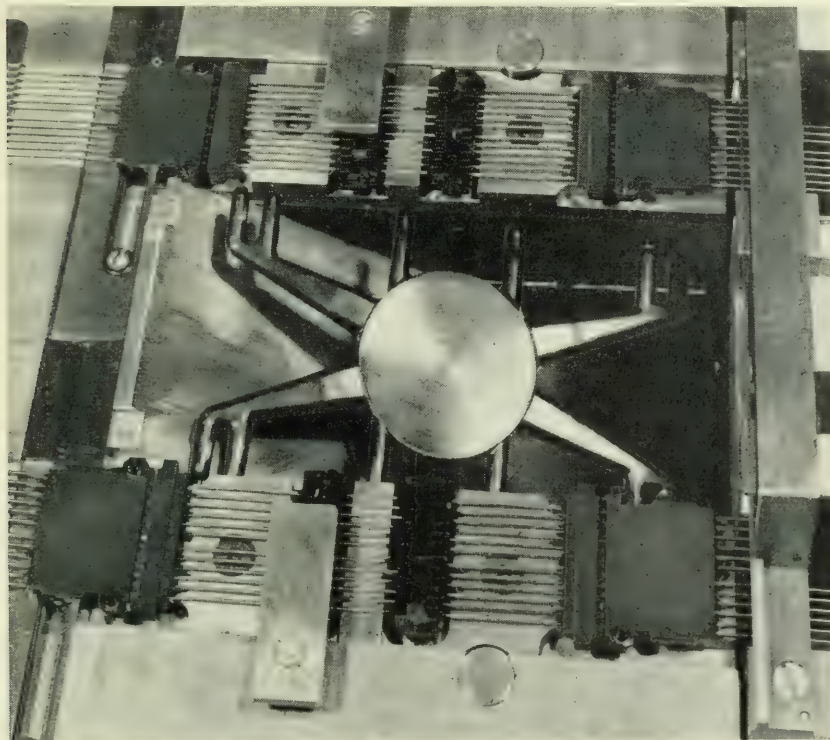


Fig. 16 — Cooling and culling section of automatic die.

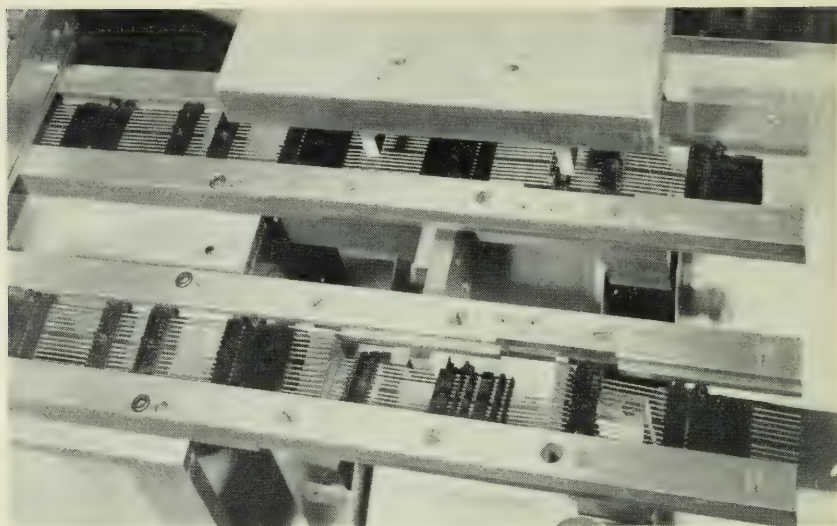


Fig. 17 — Indexing and shearing section of automatic die.

other, there is some tendency for it to warp upon cooling. To prevent this, when the die is closed the plastic embedded wire assemblage is clamped against spring loaded steel pads operating between water cooled plates and retained in this position during the next molding cycle. The closing of the die also causes the cull to be sheared off, Fig. 16. This waste material slides down a chute for removal from the press. A locating stop is built into this section of the die to establish the proper spacing for inserting a new progression of wires through the die should such action become necessary as when wire from new spools must be placed in the dies. This locating stop can be used also as a check at any time to determine whether the parts are being indexed the proper distance.

INDEXING AND SHEARING

The mechanism for indexing the continuous strands of wire through the die is located beneath and parallel to tracks built into the last die section to accommodate and guide them. It is driven by a timer controlled pneumatic cylinder to which feed heads are attached by a yoke. The feed heads reciprocate on rails beneath the guide track and transmit their motion to the ladderlike train of assemblies by a spring loaded dog which engages the resin blocks on the forward stroke. The return stroke is carried out after the press is closed.

The last operation performed by the die is that of shearing the wire in the molded assembly to form individual combs. One set of opposing shearing details for each line of molded assemblies is adjustably mounted on the base plate of the lower die half, Fig. 17. When the die is closed, the upper details butt against the upper die plate thereby forcing the cutting shears upon the wire. To reduce the force required, the cutting blades are so tapered that they cut each wire in succession.

Upon termination of the cutting operation, the parts fall free of the guide rails into chutes leading to the front of the press, thus completing the molding operation.

CONCLUSION

The original objective of embedding a multiplicity of straightened small diameter wires in phenolic resin blocks (Fig. 18) on a commercial basis has been accomplished. These wire spring relay parts are being produced at low cost to the required dimensional accuracy in automatic molding machines.

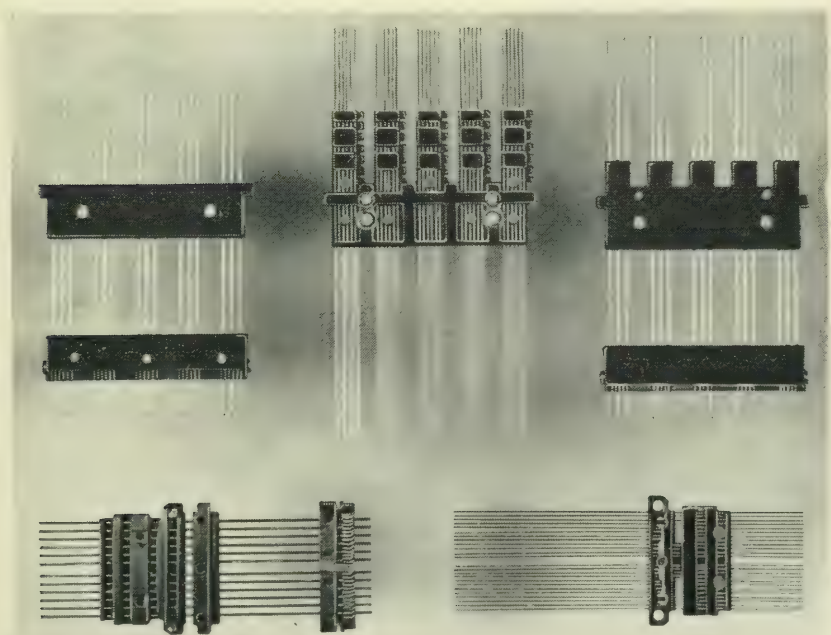


Fig. 18 — Wire block assemblies as manufactured for wire spring relays.

ACKNOWLEDGMENT

The authors wish to acknowledge the many contributions of fellow engineers to the development work which they have reported. These engineers include; in the wire straightening study, C. Paulson, A. E. Swickard, L. L. Mazza and the late N. K. Engst; and in the molding investigation, F. A. Schultz.

Some Fundamental Problems in Percussive Welding

By ERIC EDEN SUMNER

(Manuscript received February 5, 1954)

The basic processes of percussive welding are presented. Large variations in arc duration result from the spread in initiation separation, magnetic bridging effects, and the amplifying effect of evaporation. Higher voltages are shown to decrease the relative spread of initiation separation. An analysis of bridging suggests minimizing the ratio of current to separation. A welding circuit offering independence from arc-duration variations is developed. The use of a capacitive transmission line, or approximations thereto, has resulted in greatly improved process control.

INTRODUCTION

Early work in percussive welding goes back to late in the nineteenth century. Both applications and accounts in literature are relatively rare. However, this type of welding should have considerable applicability in view of some rather outstanding advantages:

1. The fact that the arc potential is approximately 15 volts permits the addition of considerable energy within a very short time and, relative to resistance welding, small currents for shorter times are possible. This allows the welding electrodes to be placed well away from the weld zone without overheating of adjacent areas. Effects of deflection due to the high electrode clamping forces can be minimized.

2. The compatibility problem between the materials and geometries of the parts to be welded are eased relative to the slower butt welding method.

3. The welds produced in a controlled process are quite strong and can well approach the intrinsic strength of the parts to be welded.

4. The percussive welding process is very fast. Use in high speed automatic production is advantageous.

The problem treated in this paper arose during a very short study program at Bell Telephone Laboratories, Inc. in connection with a new

relay development.* The paper, therefore, does not purport to be a thorough study of all phenomena of interest. The purpose is rather to enumerate the major fundamental problems and to present first order solutions. Perhaps other groups having further interest in the process will carry on basic research along the lines indicated.

BASIC PROCESS

Basically the process consists of an electrical circuit which stores energy and maintains a voltage across the two parts to be welded. This is illustrated in Fig. 1 where a wire is to be welded to a small rectangular block. A mechanical appendage, the "gun," holds one of these parts and moves towards the other, stationary part. At a separation x_0 , the arc initiation separation, the airgap breaks down and the arc is initiated. While the arc heats the opposing surfaces, forming a thin layer of molten metal on both parts, they are being brought closer together and finally come into contact, extinguishing the arc. The joint now cools and the weld is made.

A properly controlled process poses two design problems. First, it is necessary to select materials that are reasonably compatible and geometries which allow each part to be welded. The second design problem is the choice of the proper electric circuit. This paper is primarily concerned with the latter problem.

Material selection may be dictated by other considerations, but it is necessary to choose materials which are capable of producing a sound joint. Irregularities, such as gas pockets, possibly due to a low boiling point component are to be avoided. Geometries must be chosen such that in the presence of heat conduction away from the surfaces to be welded, the average temperatures of both surfaces exceed their melting points but stay below their boiling points.

A proper electric circuit for percussive welding has to supply sufficient energy to produce thin molten layers on both parts. It is undesirable to

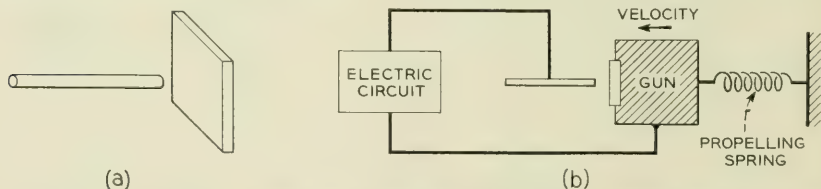


Fig. 1 (a) and (b) Percussive welding. (a) Parts to be welded. (b) Process diagram.

* A. C. Keller, A New General Purpose Relay for Telephone Switching Systems. B.S.T.J., pp. 1023-1067. November, 1952.

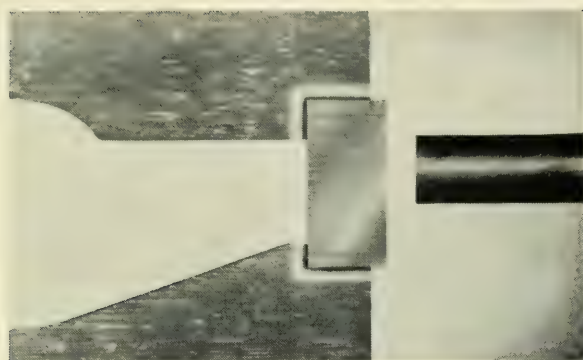
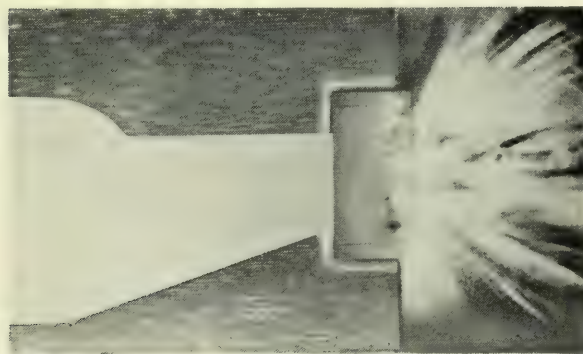
**APPROACH****ARC****FOLLOW**

Fig. 1 (c) — High-speed photographs of welding operation.

supply excessive energy because of the large amounts of material that are then burned off, and objectionable weld flash is produced. Probably the major problem is to control the energy supplied by the circuit or, indirectly, the control of arc duration.

DURATION OF ARC

In this section the variables affecting arc duration will be discussed. During the arcing period the gun moves essentially at constant speed. The arc time may then be said to be equal to the distance traveled after arc initiation divided by the gun velocity.

A. Initial Separation

The voltage at which the arc is initiated is primarily a function of the separation between the two electrodes. A series of static voltage breakdown tests was made in order to define the distribution of initiation separation under conditions to be expected in production welding of a block to a wire. The block material was 70-30 per cent cupro-nickel. The wire was 0.040-inch diameter silicon copper. Tests were taken with the wire end flat or terminated in a 60° conical point while the block surface was maintained flat. Industrial contamination as may very well be present in a production machine was simulated by the addition of a thin oil film on each of the opposing surfaces.

The results are summarized in graph form in Fig. 2. Plotted are the three σ limits* for the conditions indicated. Better arc initiation separation control is obtained with flat as compared with pointed wire ends, and clean as compared with oil contaminated wire ends. The ratio of maximum to minimum arc initiation separations to be expected is considerably lower for high voltages than low voltages. This fact alone makes operation at voltages in excess of 1,000 volts desirable.

In addition there exists an initiatory time lag† between the time that the separation reaches the static breakdown value and the moment of actual initiation. Arc duration variations due to this phenomenon are reduced by an increase in applied voltage.

B. Evaporation of Material

The arc does not cover the whole surface but is concentrated on a small area which is being heated, therefore, at a rate considerably in

* All but three out of 1,000 welds are expected to fall within these limits. It is to be noted that in view of the high reliability often required of this type of weld, conditions even further removed than three σ limits may have to be considered.

† Field Emission of Electrons in Discharges by Llewellyn, Jones and E. T. de la Perrelle, Proc. Roy. Soc. A, **216**, p. 267, 1953.

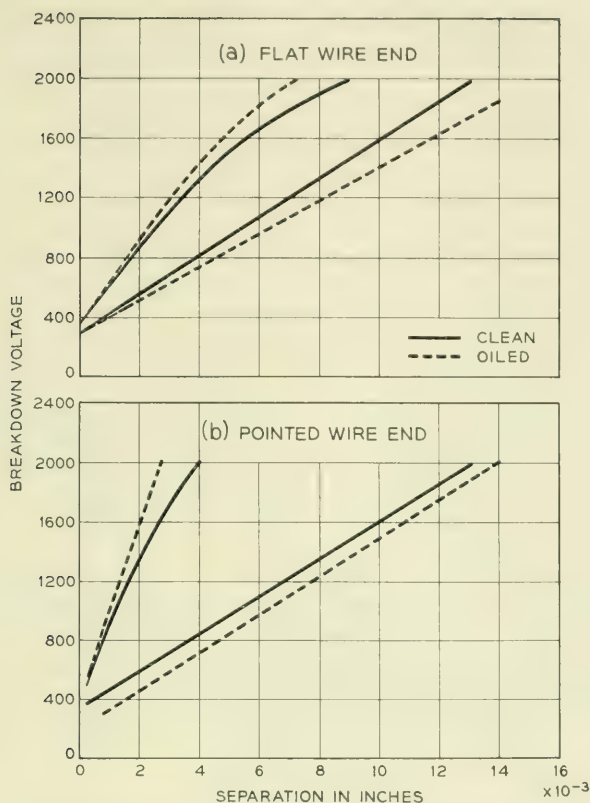


Fig. 2 — Three sigma limits of breakdown voltages between a plane surface and (a) a flat end and (b) a pointed end of a round cross-section wire.

excess of that which one would compute as average heating. As a result very high temperatures occur in the arc region and material is evaporated. This somewhat increases the arc length and the arc moves on to a new spot. As a result of this mechanism some material is burned off in the arcing period and the arc duration is therefore longer than the initiation separation divided by the approach velocity. It is quite difficult to predict the amount of material evaporated. In lieu of more extensive experiments it can be stipulated that the evaporated material should be proportional to the energy input minus the energy directly radiated to the surroundings.

It is of interest, however, to note that the phenomenon of evaporation tends to increase the spread of arc duration as computed from the initiation separation alone. This is simply due to the fact that a large initiation

separation means a longer arc duration, providing a larger energy input and hence increased evaporation. This means that the moving part has to travel further before the arc is extinguished thus further increasing the arc duration.

C. Bridging

Early experiments indicated a variation in arc duration greater than that which could be explained on the basis of variation of initiation separation and burn-off alone. It was realized that there was a third phenomenon involved. This phenomenon in which metal filaments form and extinguish the arc prematurely will be called bridging. As a demonstration of bridging the two surfaces to be welded were spaced 0.002 inches apart and voltage applied between them. The resultant arc produced a molten filament between them and a weld was formed.* With the materials used the welds produced were somewhat porous and not too strong but tests were much too fragmentary to properly evaluate this process.

Electrostatic forces are too small to account for the bridging phenomenon. A speculative explanation on the basis of magnetic forces may be attempted. Let us first examine a uniform liquid filament carrying a current I . Application of the electromagnetic stress tensor demonstrates the presence of radially compressive pressures equal to:

$$P = \frac{\mu_0 I^2 r^2}{8\pi^2 a^4}, \quad (1)$$

where

I equals total current carried by filament.

r equals radial distance from center of filament.

μ_0 equals permeability of filament material.

a equals radius of filament.

In the presence of these compressive stresses the filament will tend to elongate. Consider now the two surfaces of the parts to be welded covered with a thin film of molten material. If due to the turbulence caused by the arc a small filament forms on one surface it may tend to elongate in the presence of magnetic forces and bridge the gap between the surfaces. A detailed dynamical analysis of the formation and stability of these metal bridges is quite difficult. The following treatment is a very rough

* This may actually be an alternate method of welding with the advantage of offering excellent dimensional control.

model which will show that consideration of the magnetic forces does yield an explanation verifying the order of magnitude of the bridging observed in experiments.

The model simulates the bridging phenomenon by the translational motion of a small cylindrical filament across the gap between the two surfaces. It is argued that the magnetic energy originally stored in the arc should be comparable to the kinetic energy of the moving filament.

The magnetic energy residing within a small cylindrical filament of diameter d , length ℓ , carrying a current I is:

$$\mathcal{E}_m = \frac{\mu_0 \ell I^2}{16\pi} \quad (2)$$

If such a filament moves at constant velocity through a distance ℓ in time t its kinetic energy will be

$$\mathcal{E}_K = \frac{1}{2} M v^2 = \frac{\pi d^2 \ell^3 \rho}{8 t^2}, \quad (3)$$

where ρ is the density of the filament material. If the two energies are comparable then the bridging time is roughly:

$$t \sim \frac{\sqrt{2\pi} d \ell}{I} \cdot \sqrt{\frac{\rho}{\mu_0}} \quad (4)$$

If we assume the following as reasonable numbers:

$$\begin{aligned} I &= 1,000 \text{ amperes,} \\ \ell &= 3 \text{ mils,} \\ d &= 20 \text{ mils,} \\ \rho &= 10 \text{ gm/cm}^3, \end{aligned}$$

then equation (4) gives a transition time of 15×10^{-6} seconds. This figure is of the same order of magnitude as bridging times observed during experiments.

The rough model used here is only one stage more refined than purely dimensional analysis but seems to give reasonable agreement with experiments. Of importance is the design guide offered by equation (4). By means of proper choice of the welding circuit it is possible to select an arbitrary current versus time relationship. In order to avoid bridging effects, which are, of course, very erratic, the bridging time t should be maximized. By equation (4) the ratio of current to separation should, therefore, be minimized. Stated in words this means that if large currents are necessary for the process (this will be shown to be desirable

in a later section) they should be confined to a period when the separation between the surfaces to be welded is quite large. The current should be sharply decreased as the surfaces approach each other.

DESIGN OF IDEAL WELDING CIRCUIT

In the previous section it has been shown that arc duration will vary over a wide range. This suggests that the system be designed in such a way as to be independent of arc duration.

The procedure will be to start with the desired temperature versus time relationship of the two opposing surfaces. From this the corresponding current versus time relationship can be found and finally the circuit giving such a current distribution selected. Obviously, the "safest" temperature-time relationship is one where the temperature is kept constant at the desired level T . The corresponding current distribution will be derived on the basis of one-dimensional heat flow. Let

u = temperature

T = desired temperature at surface

a^2 = diffusivity of material

x = distance in direction of heat flow

A = cross-sectional area over which heat flow occurs

K = heat conductivity of material

V_m = voltage across arc

i = transient current

Start with the differential equation for one-dimensional heat flow:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}. \quad (5)$$

For the boundary conditions:

$$\begin{aligned} u &= 0 & t < 0 \\ u|_{x=0} &= T & t > 0 \end{aligned} \quad (6)$$

The solution is:

$$u = T \left(1 - \frac{2}{\sqrt{\pi}} \int_0^{x/2a\sqrt{t}} e^{-\beta^2} d\beta \right). \quad (7)$$

In order to determine the heat input required we must find the gradient at the surface. Expanding equation (7):

$$\frac{\partial u}{\partial x} = \frac{2T}{\sqrt{\pi}} \left[\frac{1}{2a\sqrt{t}} - \frac{x^2}{(2a\sqrt{t})^3} + \frac{x^4}{2!(2a\sqrt{t})^5} \cdots \right], \quad (8)$$

$$\left. \frac{\partial u}{\partial x} \right|_{x=0} = \frac{T}{a\sqrt{\pi t}}. \quad (9)$$

The required heat input must now be set equal to that supplied by the welding circuit:*

$$\frac{1}{2} i V_m = \frac{KAT}{a\sqrt{\pi t}}, \quad (10)$$

or

$$i = \frac{2KA}{aV_m\sqrt{\pi}} \frac{T}{\sqrt{t}}. \quad (11)$$

The current time relationship represented by equation (11) is that due to a capacitive transmission line working into a short circuit. Since the arc voltage is considerably lower than the voltage to which the line is charged, it is substantially a short circuit.

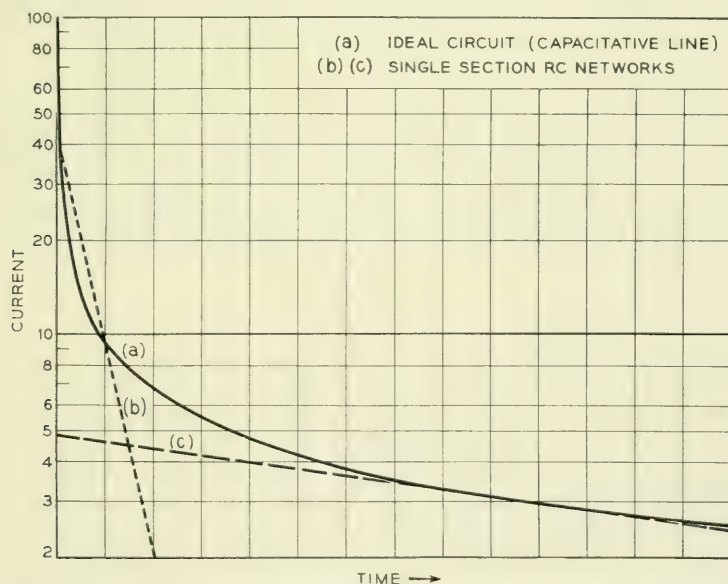


Fig. 3—Current time characteristic. (a) Ideal circuit (capacitive line). (b) and (c) single section RC networks.

* It will be assumed that the energy of the arc is divided equally by the two opposing surfaces.

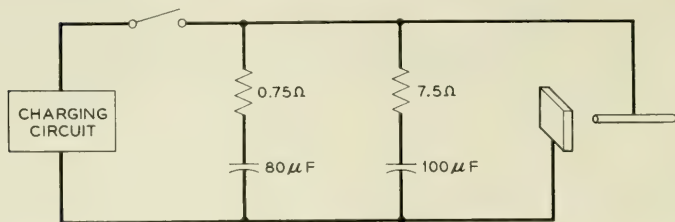


Fig. 4 — Practical welding circuit.

SELECTION OF A PRACTICAL CIRCUIT

It is probably not practical to use a distributed constant capacitive line having a current time relationship as plotted in Fig. 3. The line can, however, be approximated to the desired degree of accuracy by means of a series of r-c sections. The current discharge curve for a single r-c section plotted on the semilog graph of Fig. 2 will be a straight line. Clearly this is not a very good approximation of the ideal curve. If the constants are adjusted such that the desired initial high currents are met, the current will decay too fast allowing the surface to cool prematurely for long arc durations. If the constants are adjusted to match the desired curve for long arc times, the initial heating will be insufficient and short arc durations will produce poor welds.

A fairly good approximation of Fig. 2 can be obtained by as little as two r-c sections in parallel (Fig. 4).^{*} Use of this circuit has resulted in considerable improvement not only in the uniformity of the welds obtained but curiously enough in the control of arc duration. The reason for the latter phenomenon is that with the multiple section circuit the desired bridging characteristics can be met much more closely.

THE MECHANICAL STRUCTURE

Relatively little is known about the effect of the mechanical design of the welding apparatus on the process. Basically, the mechanical constants of interest are the mass and velocity of the gun when the arc is being extinguished and the forces propelling the gun. The gun contains kinetic energy part of which is absorbed during the impact of the two parts to be welded. The remaining part will tend to produce rebounding of the gun. Clearly the weld must have cooled sufficiently when the gun draws back such that it can withstand the forces tending to pull it apart. The time allowed for cooling is then roughly one-half the period deter-

^{*} The circuit configuration shown is equivalent to two L sections of a lumped constant line as usually depicted.

mined by the mass of the gun and the stiffness of the stationary part to be welded.

The effects on weld quality of that hammer blow produced by the gun are not well understood but there is some evidence that this blow may be advantageous in producing intimate mixing.

SUMMARY

The basic processes of percussive welding have been discussed. Large variations in arc duration are caused by the spread in the initiation separation, bridging phenomena, and the amplifying effect of evaporation.

The relative spread of initiation separation is minimized by working at high voltages, in excess of 1,000 volts. Bridging, which causes premature extinguishing of the arc, is minimized by maintaining the ratio of current to separation at a minimum.

A welding circuit offering independence from arc duration variations has been developed on the basis of one-dimensional heat flow. The analysis presented suggests a capacitative transmission line, which can however be approximated by two or more r-c sections. Greatly improved process control has been effected with this circuit.

ACKNOWLEDGMENT

The author gratefully acknowledges the assistance of J. J. Madden in all phases of experimentation connected with this project. Miss L. Mitchell performed the study of breakdown voltage. S. P. Morgan suggested the model of bridging time.

Automatic Contact Welding in Wire Spring Relay Manufacture

By A. L. QUINLAN

(Manuscript received January 19, 1954)

Welding of precious metal contacts to the new wire spring relay has presented some unusual manufacturing problems. As the name implies, the springs of these relays are wires. The contacts through which electrical circuits are established consist of small blocks of palladium accurately and securely welded to one end of these wires. The wires, arranged in a parallel array, are imbedded for part of their length in molded phenol plastic to form parts which will be designated as "combs" in this paper. There are two kinds of combs, those with the wires arranged in pairs called twin wire combs, and single wire combs. Different welding techniques are required, each of which will be described separately.

INTRODUCTION

The wire spring relay, Fig. 1, was designed with such advantages over U and Y type relays as higher operating speed, longer life, lower power consumption and lower cost, as described in a recent article in this Journal.* The lower cost will be achieved largely by reduction of assembly labor time, by reduction of adjustment effort after assembly due to greater precision in the manufacture of component parts and by extensive use of automatic manufacturing processes. To attain these goals the closest cooperation has been necessary, particularly during the design stage, between Bell Telephone Laboratories relay engineers and Western Electric development engineers. Small lots of wire spring relays of several of the early designs were manufactured by the Western to furnish the Laboratories with relays for testing. These operations provided Western development engineers with valuable manufacturing experience.

The present design of relay has been in production on a pilot plant basis. The contact welders have operated as individual units during

* A. C. Keller, A New General Purpose Relay for Telephone Switching Systems, B.S.T.J., Nov., 1952.

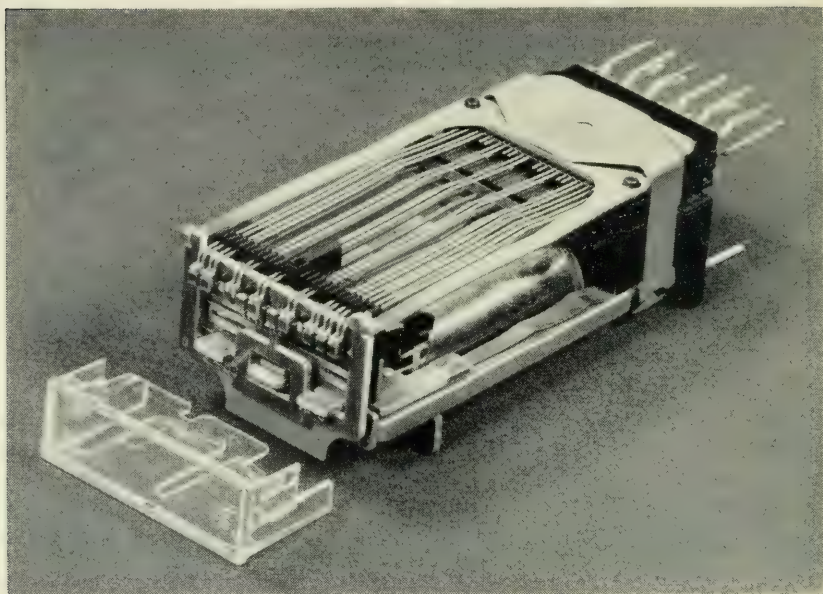


Fig. 1 — Wire spring relay.

that period as contrasted to their inclusion as components in automatic welding and wire forming lines now being placed in operation. The performance reported herein is based on individual operation.

PART I — AUTOMATIC MULTIPLE RESISTANCE WELDING OF TWIN WIRE COMBS

The problem to be solved by Western Electric development engineers was that of welding small blocks of palladium, a precious metal, to flattened ends of 0.0226-inch diameter nickel silver wires molded in twin wire combs, Fig. 2. In addition to a secure weld, close limits had to be

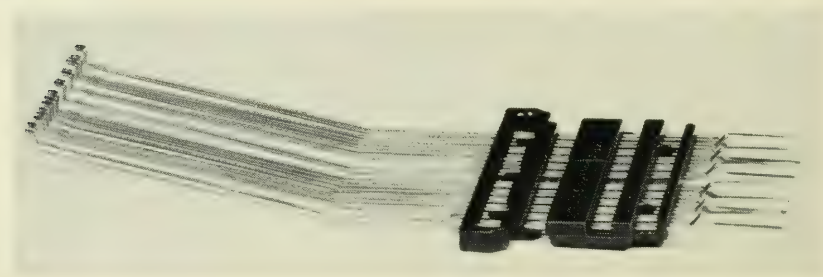


Fig. 2 — Twin wire comb.

met on the location of the precious metal. For example, Fig. 5, the upper contact surfaces had to be located with a precision of 0.004 inch. The contact itself had to be centered laterally with respect to the straight portion of the wire within 0.004 inch. Variables in the wire materials, such as elasticity, make such limits difficult to meet. The wires in the twin combs are oriented in pairs to provide bifurcated contacts, i.e., the contacts on both wires of any pair mate with one stationary contact of the single wire comb to furnish two current paths. Any number of pairs up to a maximum of twelve may be required by a code of relays. To conserve precious metal, contacts are welded only to those wires needed for a particular comb. The wires not required are clipped from the comb by a hydraulic press operated die in the automatic welding and wire forming line.

The 24-circuit capacitor discharge resistance welders, one of which is shown in Fig. 3, were designed and constructed for welding twin wire contacts. As stated before, this welder is one of the units in a welding and wire forming line. Combs from the molding operation are delivered to the beginning of the line in magazines. By fully automatic means they are removed from the magazine, carried by a reciprocating conveyor through each machine unit in the line and, when fully processed, placed in another magazine. The line of machine units, Fig. 4, forms the contact end of the wires, degreases the wires, welds the contacts to the wires, coins the contacts to the specified dimensions, forms the terminals, clips the ends to length, tension bends the wires, removes unnecessary wires and tins the terminal ends.

REASONS FOR SELECTION OF PROJECTION TYPE RESISTANCE WELDING

A capacitor discharge resistance projection welder was chosen for this job for the following reasons:

1. It is capable of welding automatically in a line of other machines.
2. It can provide the fast rate of temperature build-up required to prevent excessive heating of the small nickel silver wire ends.
3. A capacitor, charged to a fixed voltage, offers a good means of controlling weld energy within narrow limits.
4. Resistance projection welding can concentrate heat at the point required. The use of a projection at the welding interface lowers the electrical current needed to a value which can be handled by electrodes without excessive heating. By placing a projection on the metal with the lower electrical resistance, in this instance the palladium, the temperature of the palladium can be increased in the weld zone as compared to that of the nickel silver wire, thus securing a better heat balance at the joint.

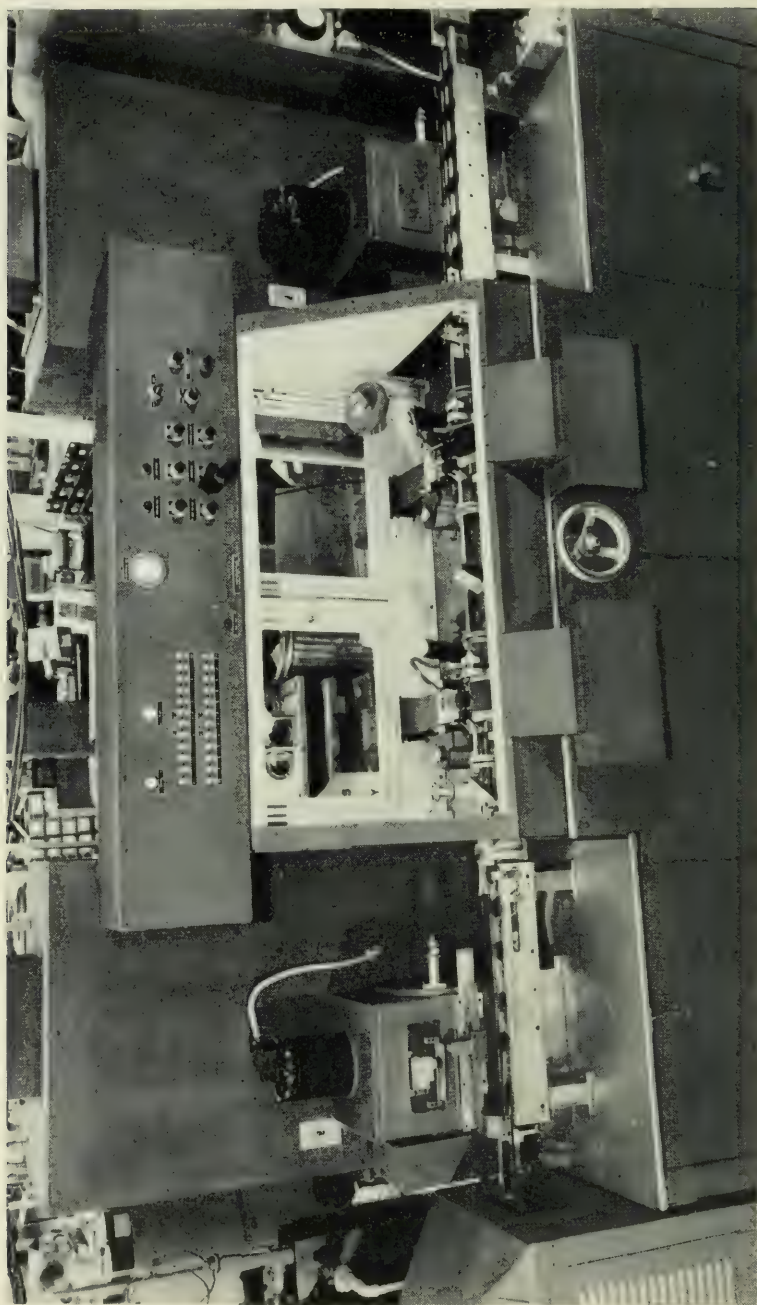


Fig. 3 — General purpose apparatus twin comb welders.

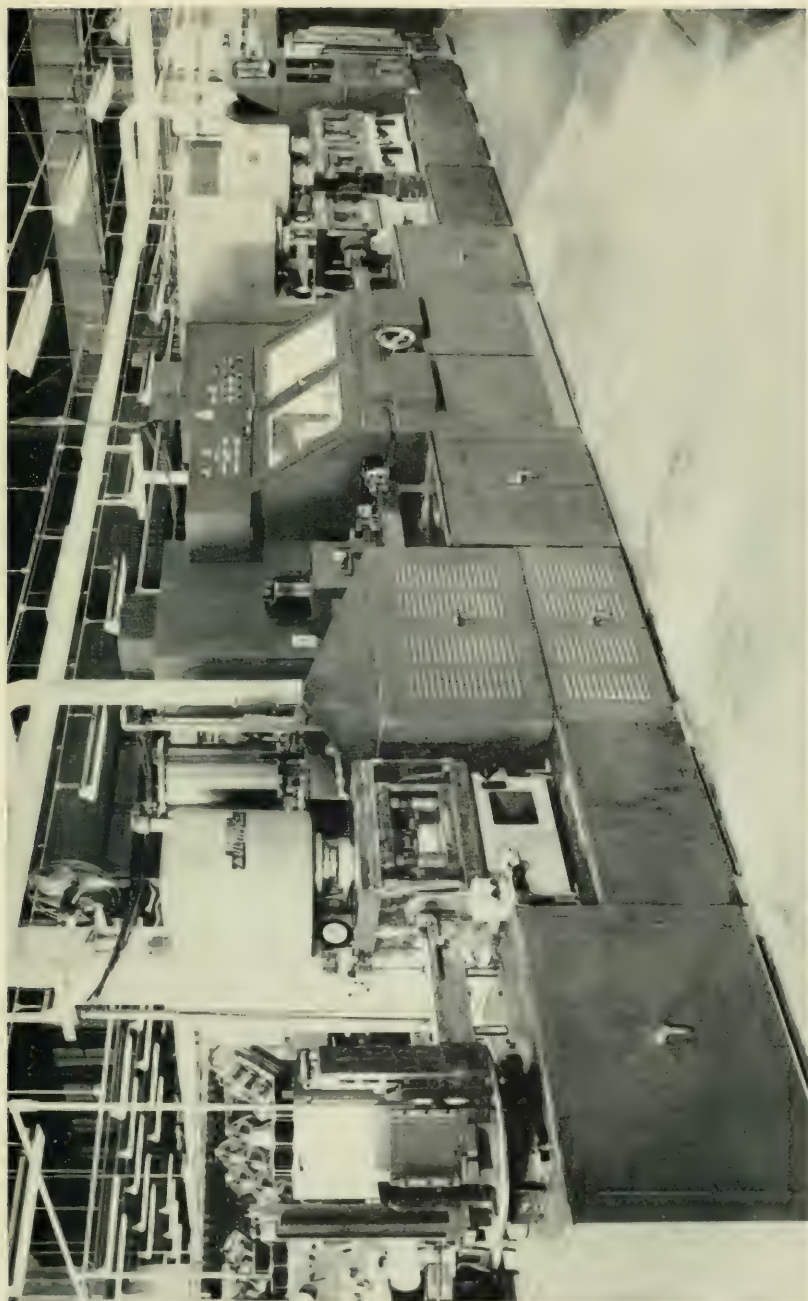


Fig. 4 Line of machine units.

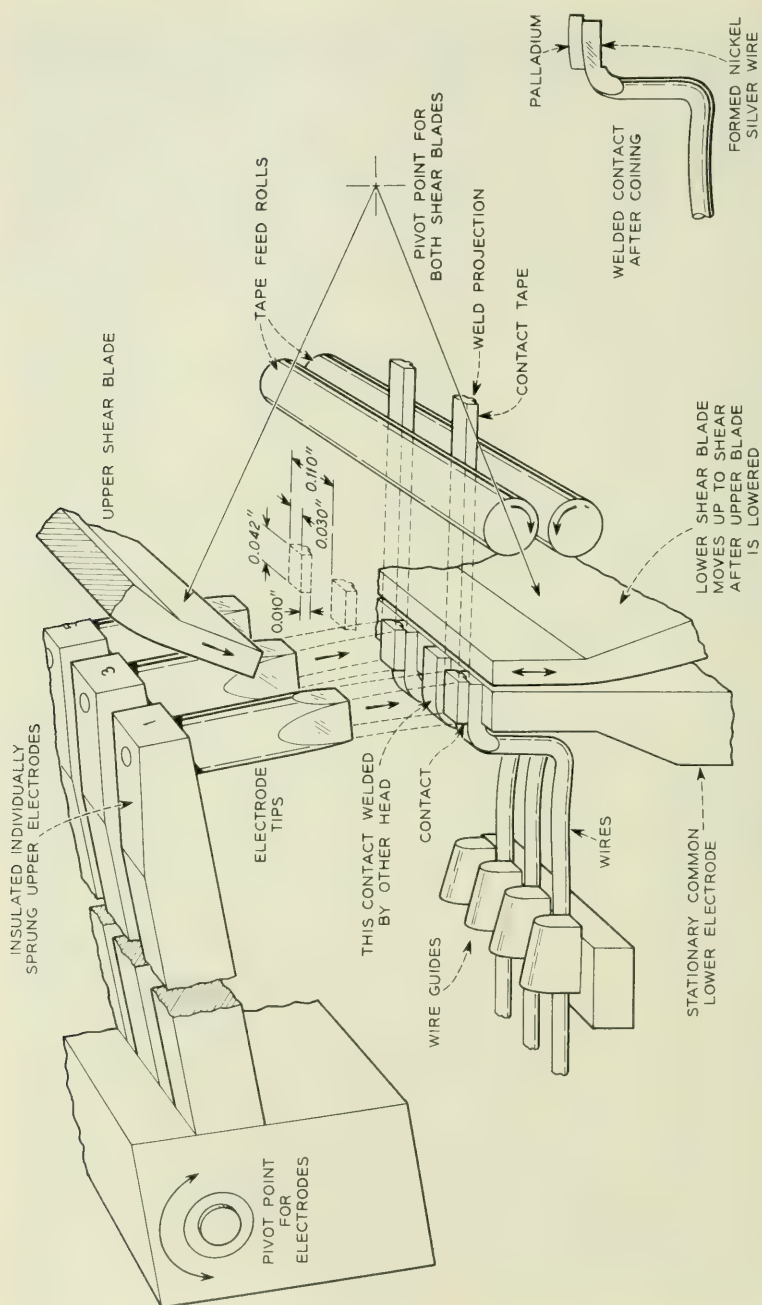


Fig. 5 — Sketch of a welding head.

The projection also confines the location of the weld to a central area of the wire end. When the weld nugget is confined to a central area, cold surrounding metal helps prevent flashing out of hot metal and results in stronger welds. To further centralize the weld nugget and to limit the projection area in contact with the nickel silver wire, the wire surface is preformed to a large radius at right angles to the weld bead.

WELDER HEAD OPERATION

Welding is done in two stages. Head No. 1 welds the odd numbered wires of each pair, after which the comb is advanced and head No. 2 welds the even number wires. This is necessary because of the small distance between the wire centers. The electrodes in each head are spaced to match the interval between odd or even numbered wire centers. Fig. 5 shows an isometric sketch of a welding head. The sequence of operations for either head is as follows:

1. The twin wire comb is advanced to a locating nest and lowered into position with the reference holes in the plastic engaged on pilot pins.
2. As the comb is lowered, the contact wires enter guide slots of a rake at a point adjacent to the plastic.
3. This rake is moved toward the ends of the wires, thereby spacing and positioning them as shown in Fig. 5. A plastic spacing member is incorporated in the relay for aligning the contacts in the relay assembly.
4. The palladium contact metal, in tape form, is advanced the proper distance to provide a contact of the specified length. These tapes, parallel with the wires, extend from guide slots for a distance slightly greater than a contact length and project over the wire ends. The tape guide slots, incorporated in a split holder, consist of steel inserts embedded in phenol fibre so designed that when the upper section is shifted laterally with respect to the lower section, the steel inserts clamp or release the tapes. Since the rake is located with reference to the tape guide, accurate spacing of contacts in the relay assembly is secured regardless of minor deviations of position of the wire ends at the welding machine. The phenol fibre mounting electrically insulates the tapes from one another to prevent part of the weld current from passing through adjacent tapes.
5. The upper electrodes, on the end of cantilever arms, are brought down on the palladium-nickel silver wire assembly.
6. The capacitors are discharged and the welds are made.
7. The upper electrodes are pivoted upward out of the way.
8. The wire guide rake is returned to a position near the plastic.
9. The upper shear blade is lowered onto the contact ends clamping them against the lower electrode.

10. The lower shear blade is moved upward, shearing off the tapes.

11. The upper shear blade is raised and the welded comb is transferred either to the second similar welder head or to the next operation.

The upper electrode tips are pins of special electrode material fixed in the ends of the supporting cantilever spring members by tapered joints, Fig. 5. Each electrode member is insulated electrically from the others and from ground by a coating of Teflon and a slotted phenol fibre guide block (not shown in Fig. 5) at the end adjacent to the electrode tips. Teflon furnishes the necessary insulation in a minimum of space and is slippery enough to allow free movement of any member. The cantilever construction makes a very light electrode assembly with the quick follow through necessary to maintain pressure on the weld area as the projection on the palladium tape is melted during the short weld cycle. A deflection of one-sixteenth inch at the tips will produce a force of about ten pounds, which is ample for this welding job. The working surfaces of the electrode tips are dressed approximately every 20,000 parts by an abrasive wheel mounted on an arbor and rotated back and forth by hand with the electrode tips pressed against the flat side of the wheel. In this manner all electrodes are dressed to a uniform length and angle simultaneously. After repeated dressings have reduced the tip length by approximately one-eighth inch, new tips are inserted.

MECHANICS OF THE WELDER

The welder is driven by an electric motor connected to the main cam shaft through a suitable reduction gear. This cam shaft is located just under the table top and extends the length of the machine. A mechanical overload clutch is provided on the output shaft of the reduction gear. All cams for the head movements are located on the main shaft. A solenoid operated clutch stops the main cam shaft when necessary. When no comb is in the weld position the weld control circuit is held open and solenoids shift either of the tape feed cams axially out of line with their followers to prevent the feeding of tape. This prevents burning the electrodes, saves precious metal and avoids loose contact pieces which might interfere with succeeding welds. A reciprocating conveyor, extending through the machine about four inches above the table top, carries the combs through the two weld positions. Excessive load, as when a part jams, will trip a microswitch and stop the machine.

TAPE SUPPLY

The contact tapes are advanced the amount required to provide the exact contact lengths by means of opposing rubber-faced feed rolls in

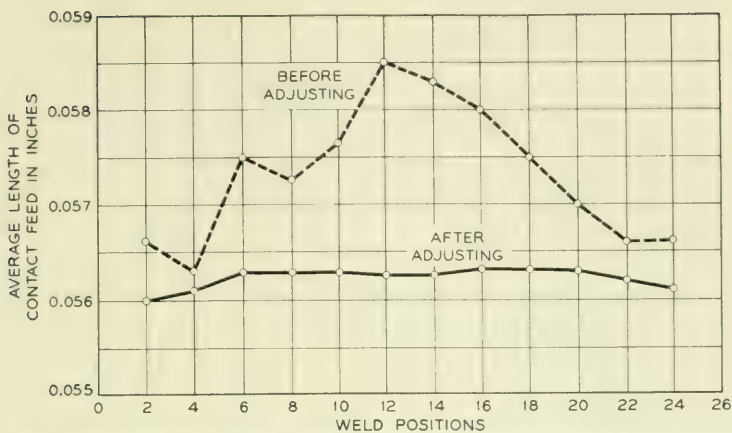


Fig. 6 — Contact length before and after adjusting hardness of rubber feed rolls and omitting gears between them.

each head. Fig. 6 shows the results of an early test on feed accuracy. Improved accuracy was obtained by providing free running tape reels. All portions of the tapes and reels are insulated from each other to prevent loss of weld energy.

ELECTRICAL FUNCTIONS

Electrically the welder has 24 separate weld circuits. Each circuit, shown schematically in Fig. 7, includes a capacitor which is charged through its own thyatron to a predetermined high voltage maintained by a voltage regulating transformer. The capacitor discharges through a thyatron and a transformer to produce the customary low voltage-high current weld energy. The transformers are located under the table near the welding heads.

The electronic equipment is mounted in cabinets to the right and left of the welder proper, as shown in Fig. 3. The capacitors are located in low cabinets, on either side, together with their solenoid controlled safety short-circuiting system. The cabinet above the welding heads contains control switches and relays. Toggle switches are provided for cutting off the weld energy for each of the 24 circuits.

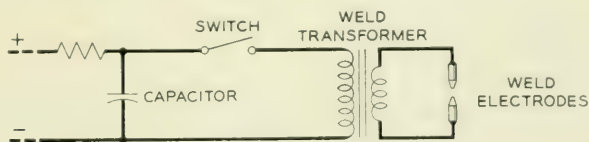


Fig. 7 — Schematic of weld circuit.

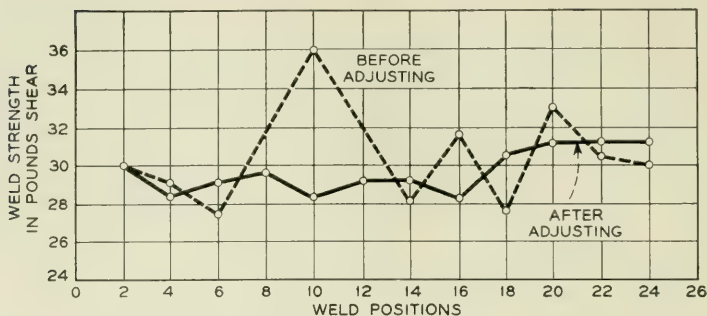


Fig. 8 — Weld strength versus resistance adjustment.

The welds occur in rapid succession in each head thereby preventing electrical interference between the many circuits involved. When combs which require less than a full complement of contacts are to be welded, the tapes in the not wanted positions are removed from the feed rolls and the toggle switches on the control cabinet are set to direct the weld energy in the sequence desired. To aid in obtaining uniform weld strength from the 24 circuits, all secondary leads from the transformers to the upper welding electrodes are of the same length. To further aid in effecting uniform weld strength, rheostats are provided in the high voltage side of the discharge circuit. Fig. 8 shows the result of adjusting these rheostats to balance the weld strengths produced by all circuits. A longer pulse would give the heat produced more time to be conducted away from the weld zone and would tend to heat more of the wire end, to the point, perhaps, of melting it completely. The three millisecond welding time has proven to be satisfactory for this job. However, the shorter the duration of the weld the higher will be the current peak required to obtain the necessary heat; the higher the current peak, the greater will be the likelihood of burning the electrodes and shortening electrode life. Adjustment of the pressure on the electrodes can be employed as a compensating factor. As the pressure is increased the tendency to burn the contacting surfaces is decreased, but the weld projection on the palladium is flattened correspondingly and the temperature of the weld, for a given current, is lowered accordingly. These variables can be adjusted to maintain optimum balance between uniformity of weld strength and electrode life.

CONCLUSIONS

More than ten million welds from this automatic machine have offered convincing proof of its capabilities. Fig. 9 indicates that good weld

strengths are being obtained. These values are shear strengths obtained on a dial indicator type of gage, reading directly in pounds, when the contact is sheared from the wire along the wire axis. A minimum test requirement of ten pounds has been established.

PART II --- AUTOMATIC PERCUSSION WELDING OF SINGLE WIRE COMBS

Percussion welding is not new. Early work goes back beyond the beginning of the century, but little application has been made of it and only a meager amount of literature is available. However, this method has a real field of usefulness as the application described in this paper will show. The original Vang process, wherein a capacitor charged to a high potential, often several thousand volts, is discharged across the gap between parts as they approach each other under a propelling force, is a good general description of the method used. The arc so produced heats the abutting surfaces before they collide so that a very thin layer of metal is brought to welding temperature. The propelling force, continuing to act, brings the parts together percussively and the weld is made. Little metal is heated and little heat penetrates the adjoining metal; therefore, the heat balance problem is greatly minimized and different metals weld together with little trouble. There is, however, the problem of protecting personnel from high voltage. Also, the two surfaces being welded must be insulated electrically from each other. This excludes the use of this process for joining the ends of the same piece of metal as in making a ring.

The project undertaken by Western Electric development engineers in the case of the single wire combs was the development of a machine for

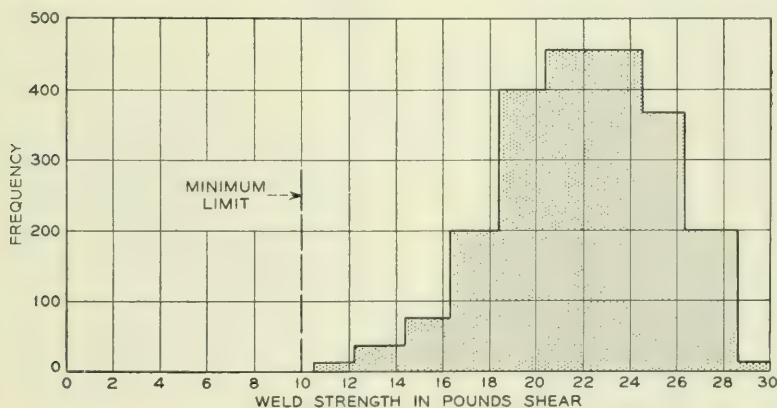


Fig. 9 — Weld strength distribution.

the automatic multiple percussion welding of contact blocks to the ends of an array of small wires extending less than a quarter of an inch from molded phenolic plastic. This array, fixed in plastic, forms a comb which is illustrated in Fig. 10. When completed this comb becomes the stationary contact member of the wire spring relay. The small blocks of metal on the ends of the wires are cut from a composite tape of which a small portion near the top and/or the bottom surface is palladium. There is a family of combs to be welded, depending on the number and type of contacts required by the code of relays into which each comb is assembled. Unlike the twin wire combs, wires which do not require contacts are left in the single wire combs, primarily to facilitate reading terminal locations during wiring into equipment. All top palladium contact surfaces must be located in the same plane across the 12 wire positions of the comb within a tolerance of ± 0.002 inch to meet design requirements. In addition, other dimensions for locating the precious metal must be held to close limits for reasons of precious metal economy.

The contact blocks for the wire comb are welded to the wire ends by the automatic percussion welder, Fig. 11, which is a unit in an automatic welding and forming line, Fig. 12, similar to but not identical with the

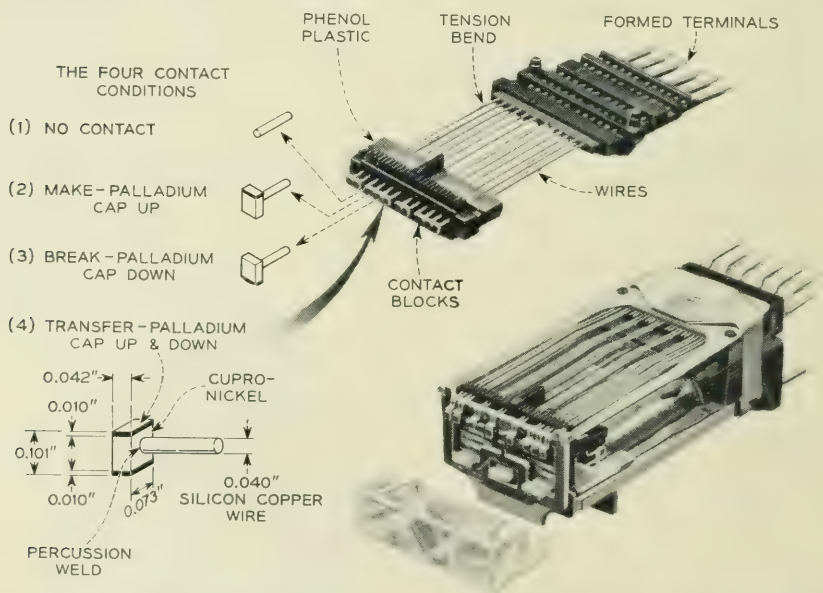


Fig. 10 — Single-wire comb with percussion welded contacts. Also view of wire spring relay.

line used for twin wire comb manufacture. These lines, by being fully automatic in operation, are interesting examples of automation.

REASONS FOR SELECTION OF PERCUSSION WELDING

Percussion welding was selected for this process for the following reasons:

1. The electrodes may be placed well away from the weld zone. The lesser current required for arc welding as compared to resistance welding makes it possible to conduct this current through the wires without heating them appreciably. The electrodes must be placed away from the wire ends so the clamping force will not deflect the wires from their normal position, thus causing a misalignment of contact surfaces.

2. A suitable heat balance in the weld zone can be obtained readily. If the slower butt welding method were used this would be more difficult because of the unequal size and differing electrical and heat conductivities of the abutting parts.

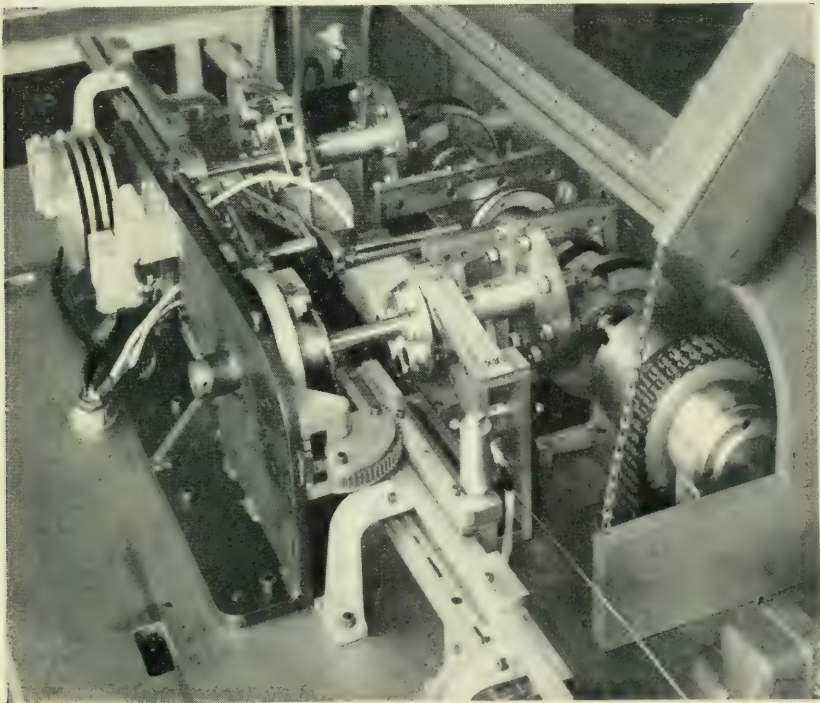


Fig. 11 — Percussion welder for contacts on single-wire combs.

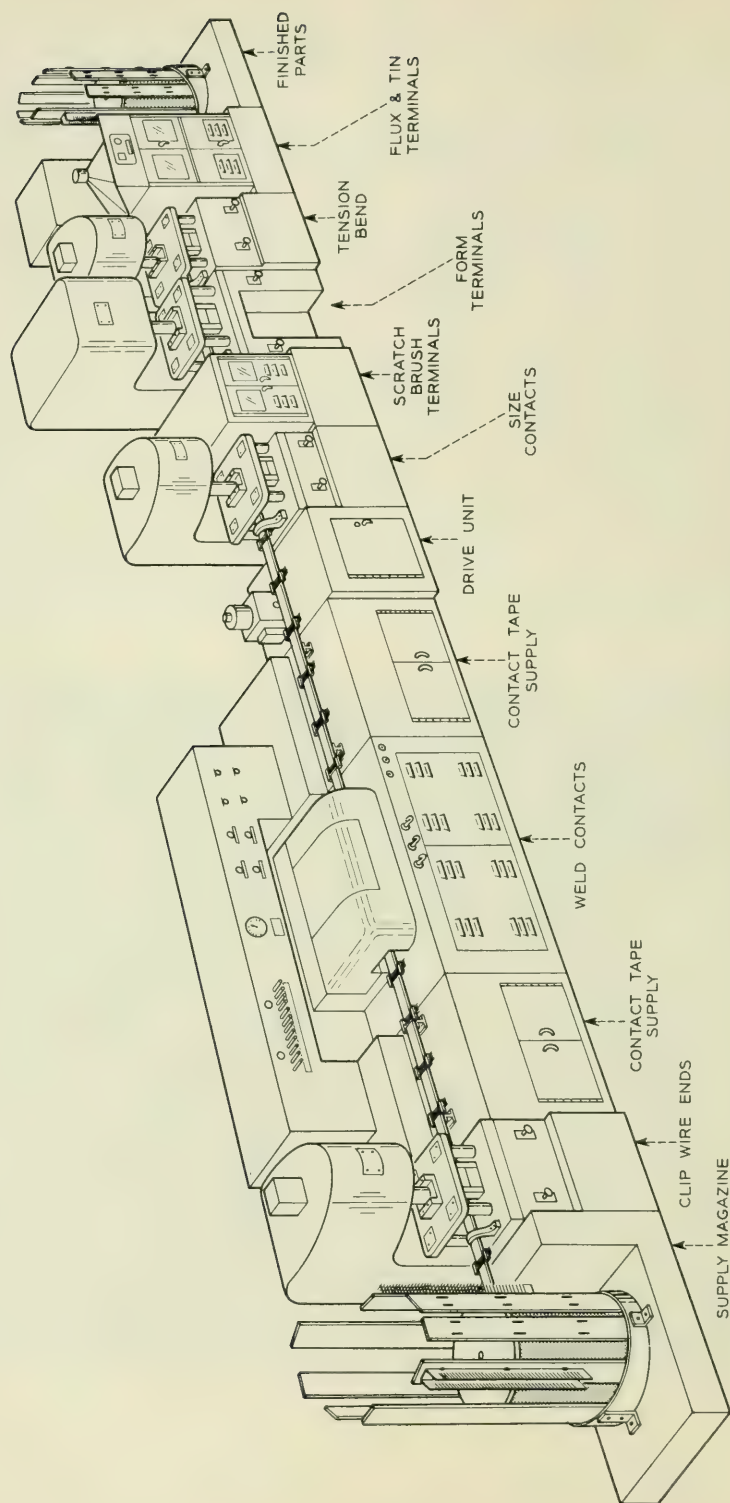


Fig. 12—Single-wire comb line.

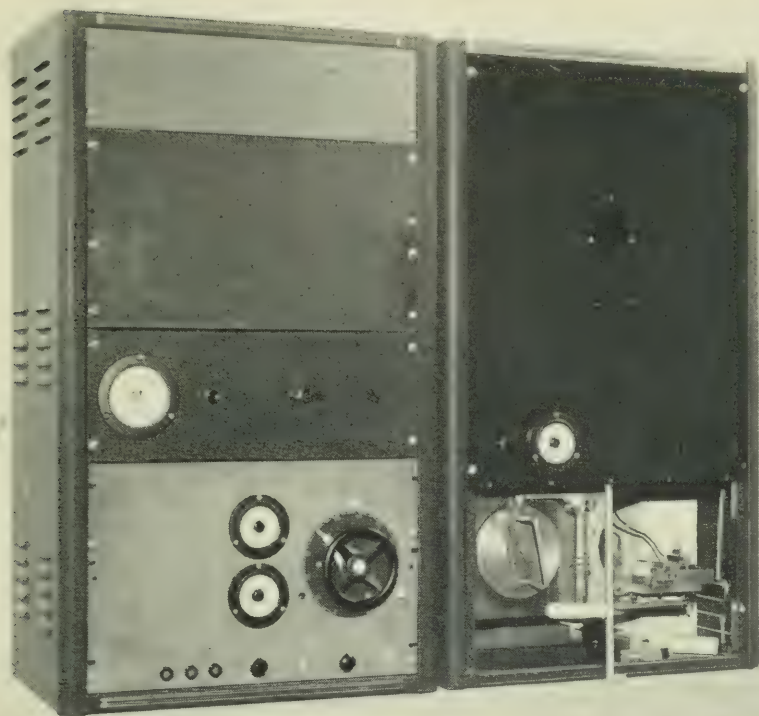
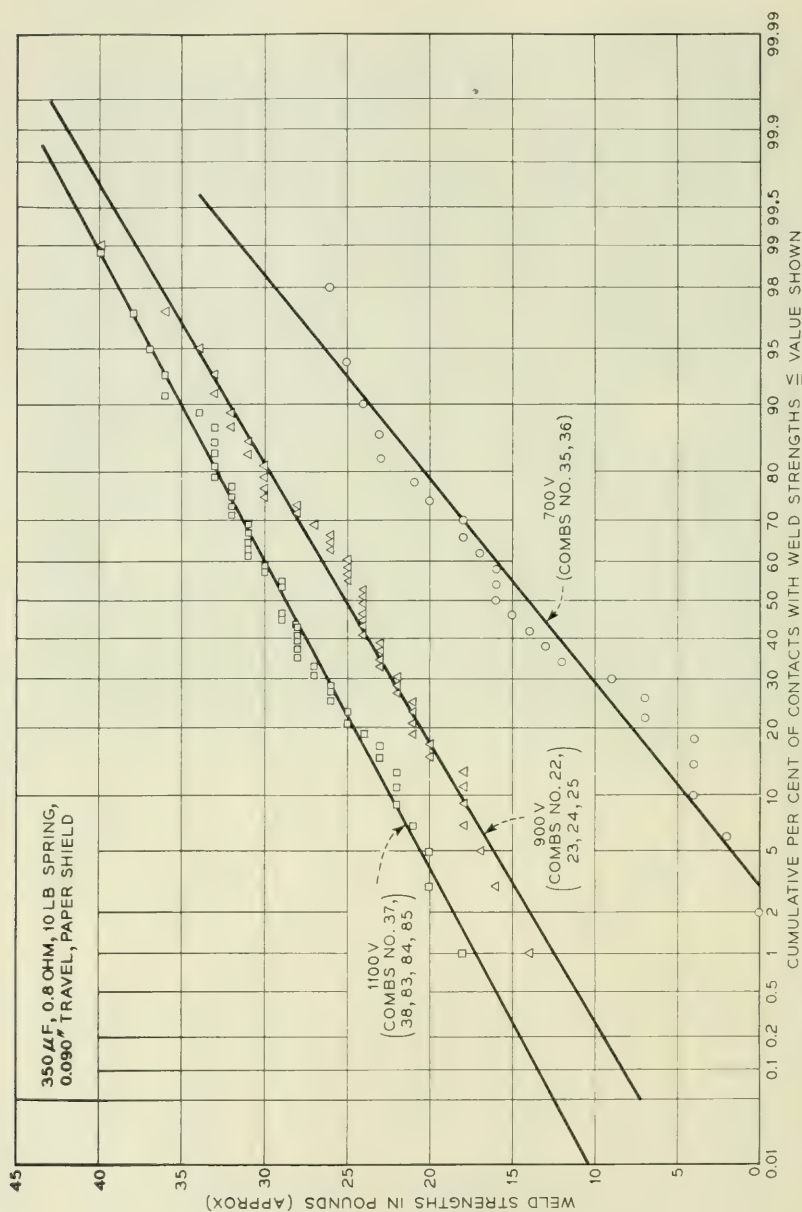


Fig. 13 — Experimental percussive welder for welding contacts to a molded comb on a "one at a time" basis.

3. The fast welding time recommends its use in high speed automatic welding machines.

EXPERIMENTAL WORK

The earliest experimental work was performed on a simple welding fixture with a spring loaded sliding jaw for retaining a contact sized piece of metal and a clamp for holding the wire. Some welds were produced but they varied widely in strength. The next fixture built, Fig. 13, was designed to weld contacts to a molded comb on a one at a time basis. A lightweight spring jawed slider held the contact and guided its travel along a fixed path. This slider was propelled by a lever actuated by a spring and controlled by a cam. The cam was powered by a variable speed drive. An extended study failed to show good weld results except at high speeds when the lever was unable to follow the cam surface and moved



Figs. 14 - Effect of voltage variation on weld strength. (Typical establishing data.)

freely, controlled only by its mass and the applied spring force. From this study the presently used spring actuated welding gun has evolved.

To determine the operating conditions that would be required, an investigation was made in which a group of 48 contacts was welded under carefully controlled conditions following which one condition was varied and the tests repeated. Many curves were established in this way for such variable factors as voltage, capacitance, resistance, spring force, and distance traveled; all related to weld strength. Weld strength was measured by a hook-pull gage developed at Bell Telephone Laboratories. The "burn-off" i.e., reduction in length of the wire and contact as a result of the arc heat, was determined in many of these tests because it usually provided a measure of weld strength uniformity. Many other conditions were investigated such as the weldability of different metals; the effect of cleanliness and of contamination on the joining metals; the influence of the physical form of the joining metals, such as pointed, rounded, or flat surfaces on wire ends; the effect of atmospheric conditions, of the presence of various gases, or a stream of compressed air, and of shields in or near the weld zone. The nature of the weld flash deposit was studied. Neither streamers of base metal which might extend over the edges of the palladium caps nor loosely adhering and easily dislodged metallic particles could be tolerated. Charted data from some of these tests are shown in Figs. 14, 15, and 16.

Another phase of the investigation was concerned with obtaining such

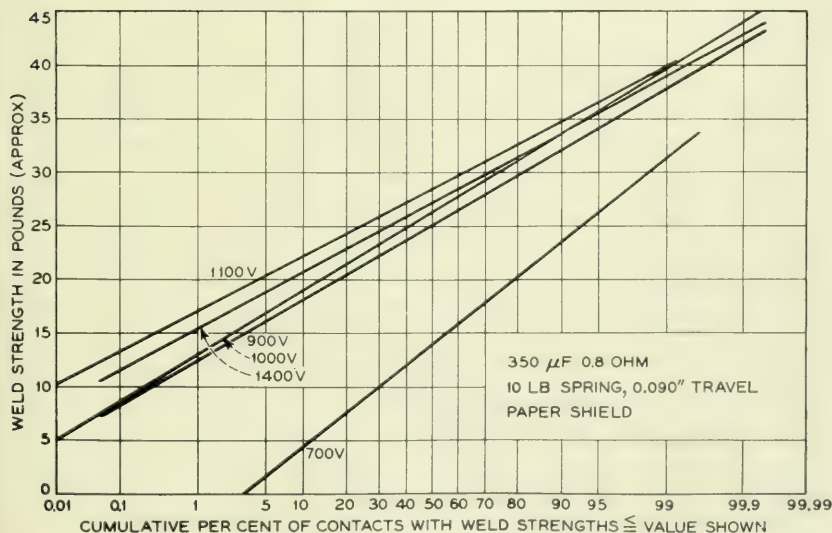


Fig. 15 — Effect of voltage variation on weld strength.

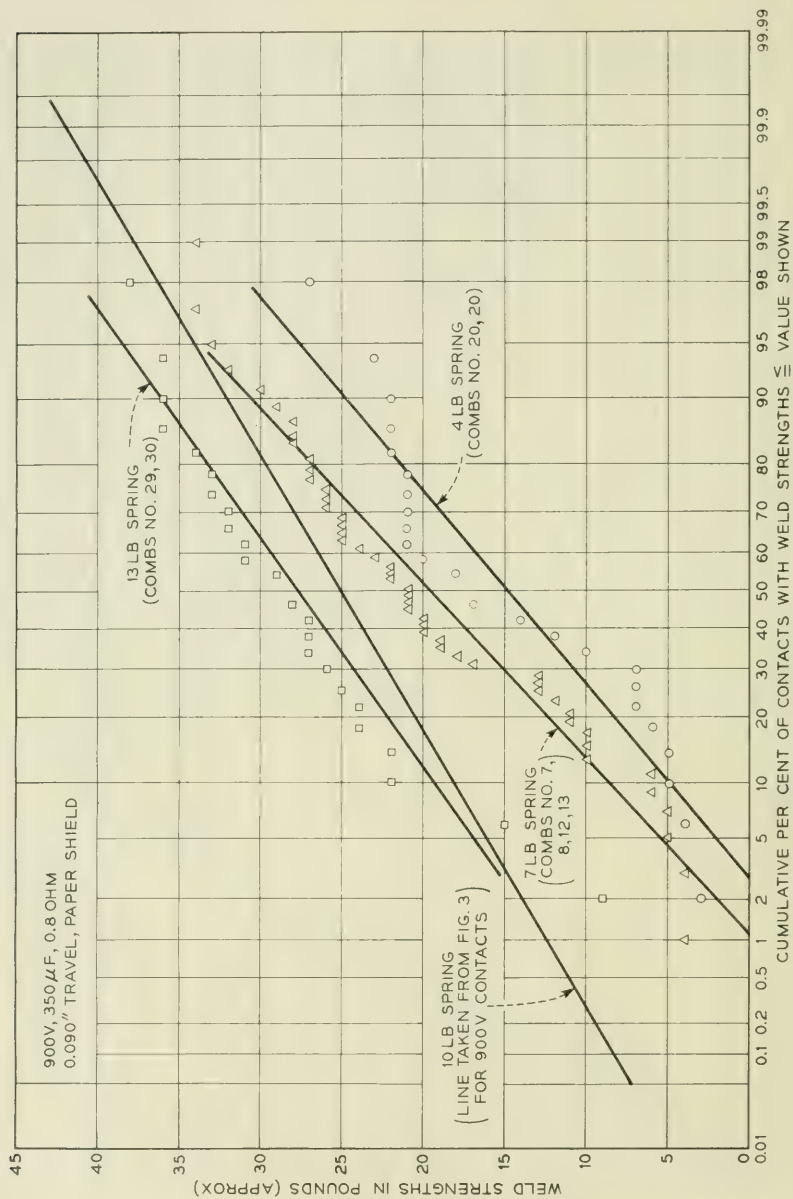


Fig. 16 — Effect of spring pressure on weld strength.

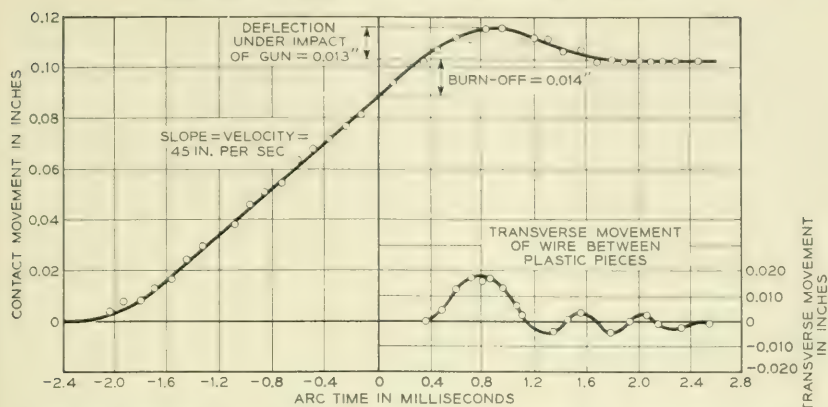


Fig. 17 — Motion of contact during percussion welding plotted from measurements on film.

evidence of what occurred during the welding operation as could be revealed by high speed motion pictures. Fig. 17 shows the kind of information obtained from these films; notably the speed of approach, the amount of burn-off, and the deflection of the comb wires back of the front plastic block upon impact by the welding gun.

In connection with a study directed toward a better understanding of the mechanism of percussion welding undertaken by E. E. Sumner of Bell Telephone Laboratories tests were made with an electro-capacitive transducer, oscilloscope, and polaroid camera arrangement for the purpose of recording variations in the velocity of the welding gun. The characteristics of the current discharge during welding was recorded by another oscilloscope-camera setup. Burn-off was measured and broken weld surfaces on the contacts were examined and photographed. These tests pointed to the value of a parallel capacitor discharge circuit as a solution to the arc duration variability problem. A commercial trial of parallel capacitor circuits at Western Electric demonstrated the merit of this type of circuit, which will be described in more detail later. Calculations from the transducer traces indicated that the striking force of the contact and contact holder upon impact with the comb wire is less than seventy-five pounds. The welding gun was operated on a reed mounting during Mr. Sumner's study.

Early experimental work on percussion welding showed that small gas pockets in the weld zone were causing weak welds. Nickel silver was used at that time. Its component of easily volatilized zinc was suspected of causing the trouble. Then various metals and alloys were tested and silicon-copper was chosen for the wires in the comb. This alloy,

however, was not suited for use as the base metal of the contact block because it was difficult to weld in the roll welding process used for fabrication of the contact tape. A 70-30 per cent cupro-nickel was selected for this base metal because it approximated the resistance of the palladium, a necessary condition for roll welding, and it did not have an easily volatile component to weaken the percussion welds.

Another subject investigated was the speed of approach of the parts during the percussion welding operation. The comb and its array of

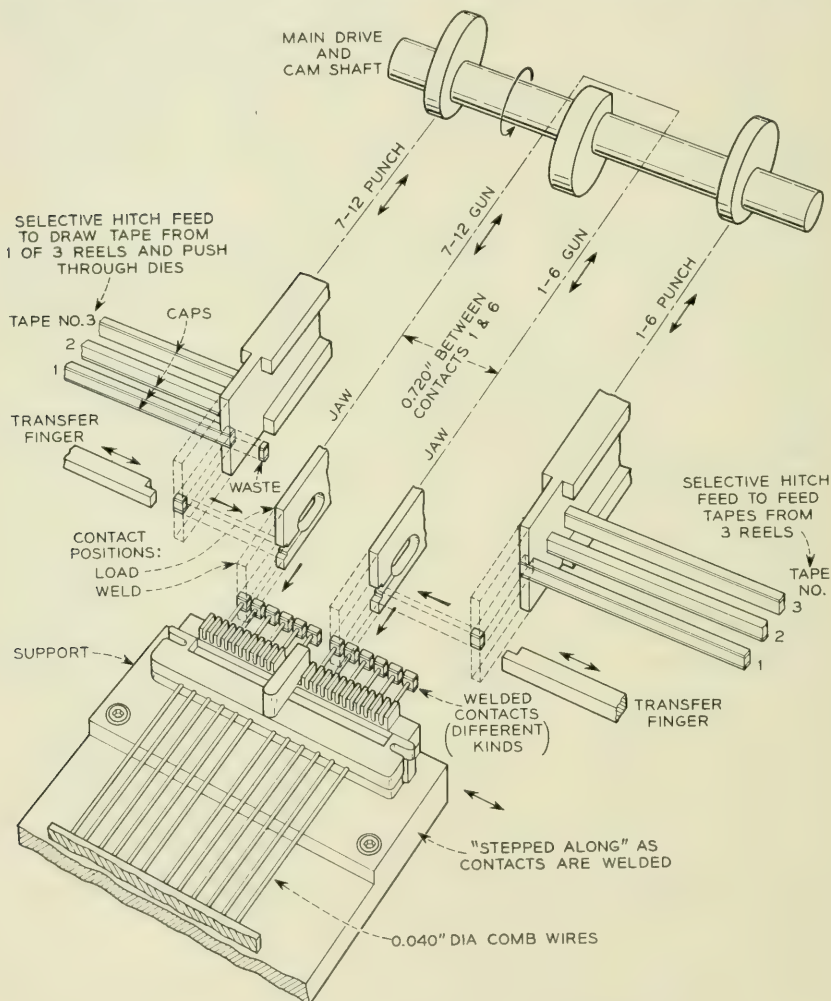


Fig. 18 — Schematic of the percussion welder.

wires is held stationary and the contact block is moved to the wires to make the weld. The speed of approach is an important factor in controlling the arc duration and the impact force. Various speeds of the contact block carriage from approximately 1 to 80 inches per second were tried. The best welds were obtained by a spring actuated gun which attained a velocity at impact of approximately 40 inches per second.

AUTOMATIC WELDER OPERATION

The automatic percussion welder contains duplicate welder heads which are mirror images of each other. Two contacts are welded at a time, one on each half of the comb, Fig. 18. After each welding operation or one cycle, the comb is indexed to the next pair of wire centers. Six cycles complete the welding of twelve contacts, or a lesser number if required. The guns do not weld at exactly the same time. There is an interval of one degree of revolution of the main shaft between the firing points to prevent electrical or mechanical interference. The welder was designed to select the type of contact, cut it from the tape and weld it in one cycle, to avoid the handling problems associated with precutting and magazing contact blocks.

TAPE SUPPLY AND SELECTOR

One of four contact conditions will apply for each wire; (1) no contact, (2) contact with palladium cap up, (3) palladium cap down, or (4) palladium cap both up and down. This requires three reels of tape on the right for one head and three on the left for the other. Adjustable knobs on both right and left tape feed cams are set for one of the four tape feed conditions for each wire position of the comb. Thus, any combination of contact conditions can be set up to make parts for any code of relay.

CONTACT SHEAR AND TRANSFER

The three tapes enter the shearing die through individual openings. However, only that tape selected by the tape selector is fed into the die and subsequently sheared by one punch stroke, Fig. 18. The tape is punched from such a direction that the base metal is not dragged over the boundary line into the palladium zone. This avoids contamination of the palladium. As the contact is blanked out, the walls of a notch in the punch confine it on the precious metal sides to prevent distortion. The punch delivers the contact to a transfer position at the end of the shearing stroke. There a transfer finger pushes the contact out of the punch notch, through a guide channel and into the waiting gun jaws.

WELDING GUNS

The welding gun is a light reciprocating member that carries two opposing steel fingers or jaws to receive the contact block from the transfer finger mentioned above. The jaw opening is a few thousandths of an inch less than the nominal height of the contact, however, the edges are beveled so that when a contact is pressed against them they spring open, the contact enters and is held securely in place. After welding the jaws are pulled off the contact. At the extreme return travel of the gun any contact which might remain in the jaws because it was not properly welded is removed by an ejector blade. When in the loading position, a portion of the blade stops the travel of the contact through the jaw opening so it is held in a uniform position and will be located on the wire with precision.

GUN MASS CONSIDERATION

During welding the comb must be supported accurately to meet the close contact location requirements. It must be supported securely to withstand the impact of the guns, or weak welds may result. The mass of the gun is important. Evidence indicates that more uniform and higher weld strengths are obtained with lightweight guns. A magnesium gun weighing about 60 grams produced better welds than did the original steel gun weighing about 130 grams. A newly installed steel gun weighing about 30 grams appears to be even more satisfactory. Other guns are being designed and tested to further check various features. A striking force of approximately 75 pounds tends to loosen the wires in the plastic and to produce weak welds. The 60 and 30 gram guns propelled at about 40 inches per second at impact produce less than half this force. The velocity during the arcing period is important to control the amount of heating. One-half cycle of vibration of jaw and wire after impact is the time available for the weld to freeze before a tension strain is placed on it. This time has been measured in the laboratory by the use of a transducer. Weak welds resulted when the time was less than 1 millisecond.

ELECTRICAL FUNCTIONS

Fundamentally, each weld circuit includes a capacitor which is charged during a small portion of each cycle and subsequently discharged through a resistance in series with the weld. During the charging period, which is controlled by a cam and microswitch, the contact and the wire end of the comb are separated electrically at the weld point. A mul-

tiple leaf brush under considerable pressure connects the one side of the circuit to the terminal end of the individual wire being welded. The gun, and through it the contact block, is connected to the other side of the circuit. After a cam frees the gun, the spring propells it toward the wire end and an electrical arc is established by the high potential (900 to

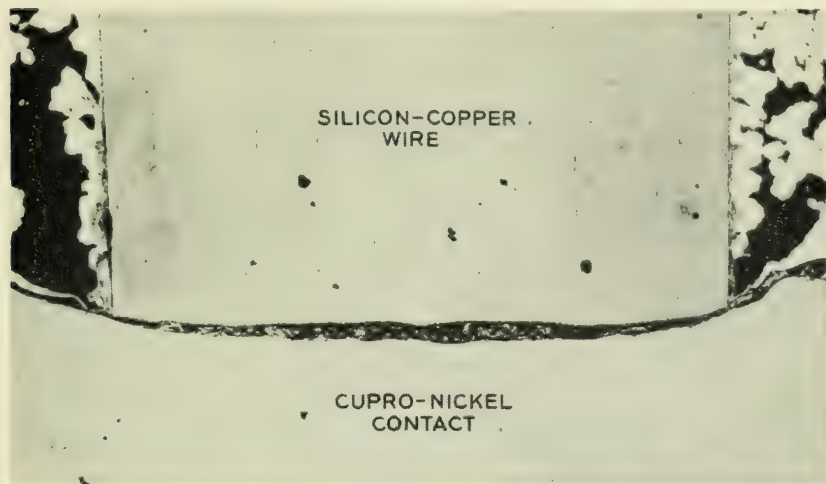


Fig. 19 Section of a percussion weld. Original amplification 100 $\times$.

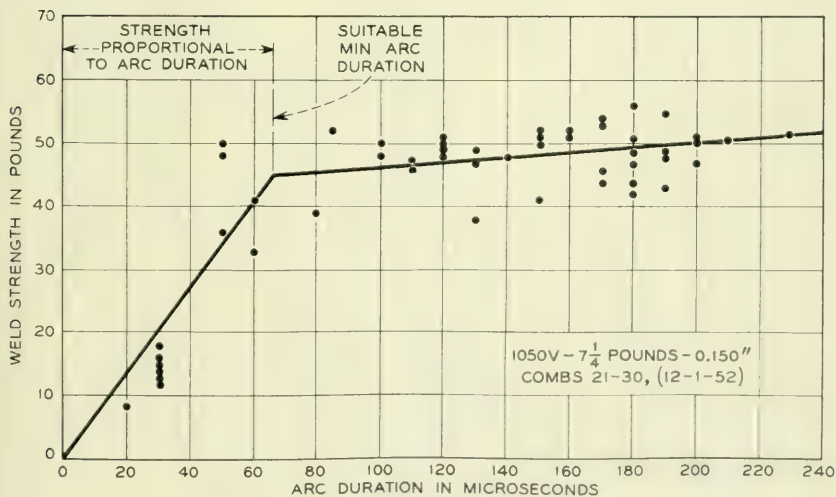


Fig. 20 — Weld strength versus arc duration.

1800 volts) just before the parts touch. The arc initiates itself when the gap has been reduced to a few thousandths of an inch. A portion of the abutting surfaces of both the wire and the contact base metal (normally 0.005 inch to 0.010 inch) are melted and expelled in liquid and gaseous states before the molten surfaces are forced together. The arc is extinguished when it can no longer melt and expel metal to maintain a gap. Under good operating conditions it persists from 0.1 to 0.4 milli-seconds. Nearly all of the heated metal is expelled from the joint during the welding operation as illustrated by Fig. 19. This micrograph of a typical sectioned percussion weld shows only a 0.001 inch to 0.002 inch thick layer which was melted or heated sufficiently to change the structure.

A small stream of compressed air is directed into the weld area to remove gaseous, possibly ionized, arc products so that they will not

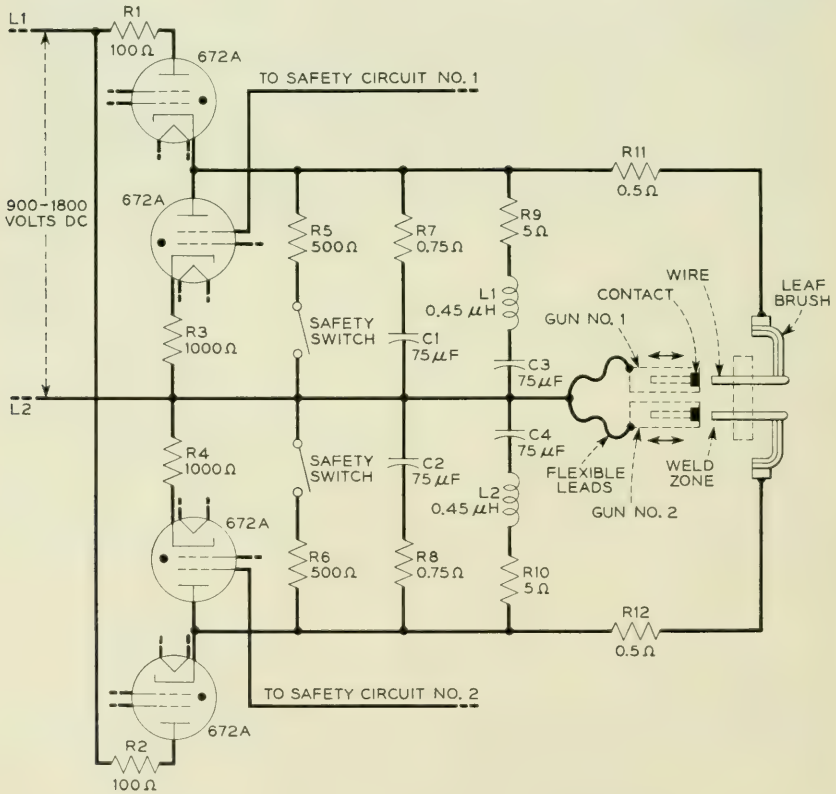


Fig. 21 — Schematic of percussion welder circuit.

interfere with the initiation of the arc during the next weld cycle. Limited tests made with nitrogen and helium atmospheres gave no indication of improvement in weld quality.

In the course of the many studies employing different kinds of instrumentation, an electronic counter was used to measure successive arc durations. Variations from 20 to 230 microseconds were observed for the prevailing conditions during early tests. The contacts welded were measured for strength of weld and a correlation found between short arc durations (up to 65 microseconds) and weak welds, Fig. 20.

Circuit impedance was found to be important in producing uniform weld strength. The assistance provided by the Laboratories in the early work led to better control of arc duration. Hawthorne continued this study and arrived at the circuit, Fig. 21 which greatly improved the uniformity of welds. The improved circuit resulted in less arcing in the jaws, thus prolonging their life and precision. Uniformity of burn-off and of weld strength both showed marked improvement. Fig. 22 shows a typical weld strength distribution for this circuit based on standard

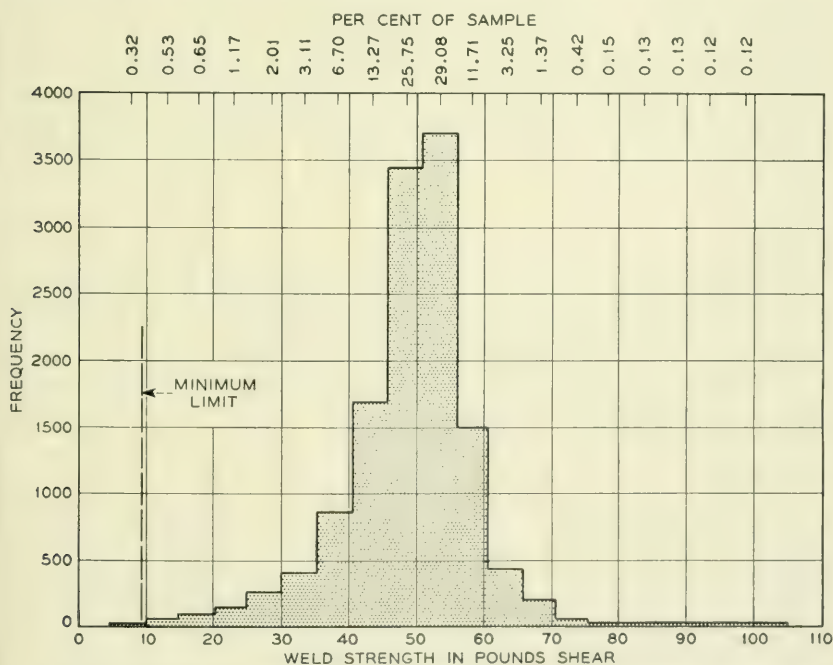


Fig. 22 — Weld strength distribution for 13,200 samples tested from 600,000 welds made in production.

shear test data. In a limited number of samples which were tension tested failure occurred more frequently in the wire than in the weld.

JAW LIFE

Jaw life is dependent upon many factors, but one of importance is the prevention of accidental conditions which establish an arc directly to the jaws. The jaws are adjusted so they will not touch the wire end and thereby discharge the capacitors if no contact is in the jaws. However, if only a very short length of contact metal is cut off and put in the jaws it may start the arc but there may not be sufficient material or the material may not be held securely enough to keep the arc from burning the jaw surfaces. Another possible trouble condition occurs when the end of the wire is misplaced so it touches a jaw. A safety circuit is provided with microswitches which trigger thyratrons to discharge the capacitors when either of these conditions occur. Signal lamps are provided to indicate at which microswitch the trouble is occurring. A reset microswitch is used in the test for shorted contacts so the indication can be held to a later time in the cycle. At the end of the stroke this switch is reset. After the part is nested approximately a quarter of a second is available to test for safe welding conditions. A cam-actuated microswitch controls the time of test and, if conditions are at fault, the weld energy is discharged through a thyatron before the parts are brought together for welding.

SAFETY

Safety for personnel from high voltage is provided by door switches, solenoid released shorting bars and bleeder resistors on the capacitors. Safety from mechanical jams is provided by a slip clutch on the main drive gear and by a pull out clutch and automatic stop switch located in the comb transfer drive mechanism.

CONCLUSIONS

After making millions of welds by automatic percussive welding methods, it is found to be a method well suited to this job. Accuracy of location and good weld strength are obtained. This welding method is especially useful where speed and precision are desired, and where joints must be made between dissimilar metals. Metals of high heat conductivity and high electrical conductivity join readily by this method.

ACKNOWLEDGMENTS

It is not possible to give due credit to all engineers who contributed to the development of these welding machines and to the preparation of this paper. Particular mention, however, will be made of H. Beedy, D. R. Brown, R. L. Kessel, G. A. Mitchell, J. M. Roach, J. P. Seider, R. Spillar, and the late Dr. B. J. Babbitt who, as Western Electric development engineers, engaged in various phases of this investigation. In addition to acknowledging the part played by numerous Bell Laboratories engineers, I desire especially to express my appreciation to D. C. Koehler and J. J. Madden for their able assistance while loaned to Western Electric during the early stages of the percussion welding study.

Electronic Relay Tester

By T. E. DAVIS and A. L. BLAHA

(Manuscript received September 24, 1953)

An electronic relay tester has been developed to gauge the contacts of a relay while it is pulsing. The device provides a visual presentation of the gauging position of up to 16 contacts at one time. The equipment has been developed primarily as a means for rapid inspection and concurrent adjustment (when needed) in relay manufacture. It may also be used as a laboratory instrument for studies of timing, chatter, and other performance characteristics of relays.

INTRODUCTION

Electronic equipment has been developed to gauge up to 16 relay contacts simultaneously as the relay operates or releases in pulsing. The system provides a visual presentation on an oscilloscope of the gauge points where the relay armature opens or closes the contact. The position of the armature is shown by the horizontal position of the scope beam. Each particular contact is represented on the scope by one of 16 horizontal lines generated by the motion of the armature. A vertical step on a line indicates the gauge point at which the contact operates.

The device consists of three main parts: (1) An electrostatic gauge to indicate the position of the relay armature, (2) An electronic scanner that continually switches across all the contact pairs, and (3) A brightness control circuit to intensify the beam of the scope when the armature is moving. Fig. 1 shows a schematic diagram of this equipment.

As shown on Fig. 2 the scope, test relay jack and electrostatic gauge are mounted on a bench. The scanner, power supplies and control circuits are mounted underneath the bench.

ELECTROSTATIC GAUGE

As the relay armature moves, its position is indicated at all times by the horizontal position of the scope beam in response to the output voltage of the electrostatic gauge, which is applied to the horizontal plates.

This output voltage varies with the position of the armature which serves as one plate of an air capacitor controlling the frequency of an oscillator. The voltage across a coupling impedance in the plate circuit of the oscillator is rectified to provide the output voltage of the gauge.

The gauge which is shown schematically on Fig. 3 may be divided functionally into three main parts: (1) An oscillator which operates in the region of 100 megacycles; (2) A detector sensitive to the frequency of the oscillator and consequently also sensitive to the position of the armature; and (3) A "zero set" cathode follower which connects the detector output to the input of a dc amplifier. The dc amplifier has suffi-

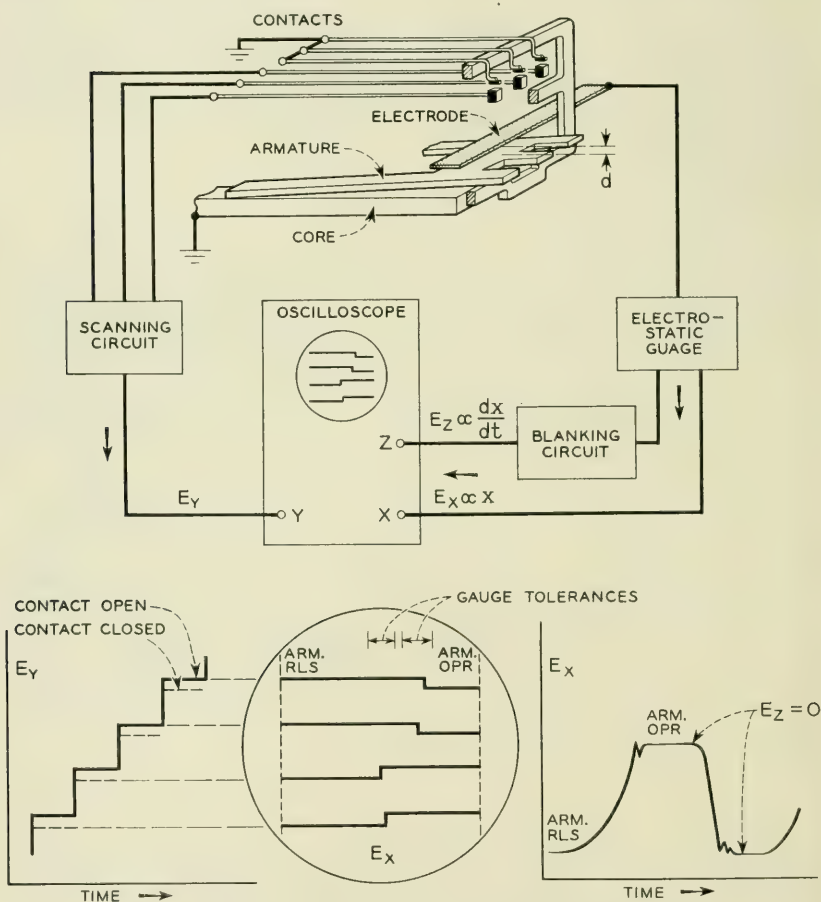


Fig. 1 — Schematic diagram of electronic relay tester.

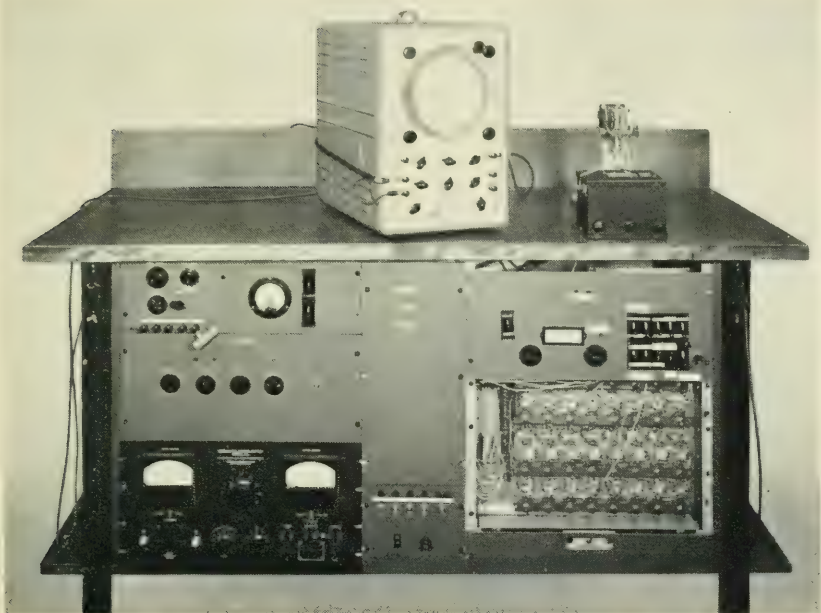


Fig. 2 — Photograph showing arrangement of equipment.

cient gain to provide adequate deflection voltage to the horizontal plates of the scope.

The oscillator comprises a 6AU6 pentode tube and an associated tank circuit. The tube is connected as an electron coupled oscillator with the energy for the oscillations being supplied by the screen circuit. The plate output is a modulated current that passes through the screen and suppressor grids. Considerable isolation of the coupling circuit from the oscillator is obtained by this arrangement.

An antiresonance circuit similar to the tank circuit is used to couple the plate of the oscillator tube to the rectifier. Since the electron coupled oscillator acts as a high impedance generator or constant current source the voltage across the coupling impedance is proportional to the value of the impedance. Hence as the oscillator frequency is shifted by a change in the position of the armature a corresponding shift occurs in the ac voltage across the coupling impedance and rectifier tube (6AL5). The output voltage of the rectifier also shifts with the input to the rectifier since the changes are slow compared with the time constant of the output circuit of the rectifier.

In normal operation the output of the rectifier may vary from 6 volts

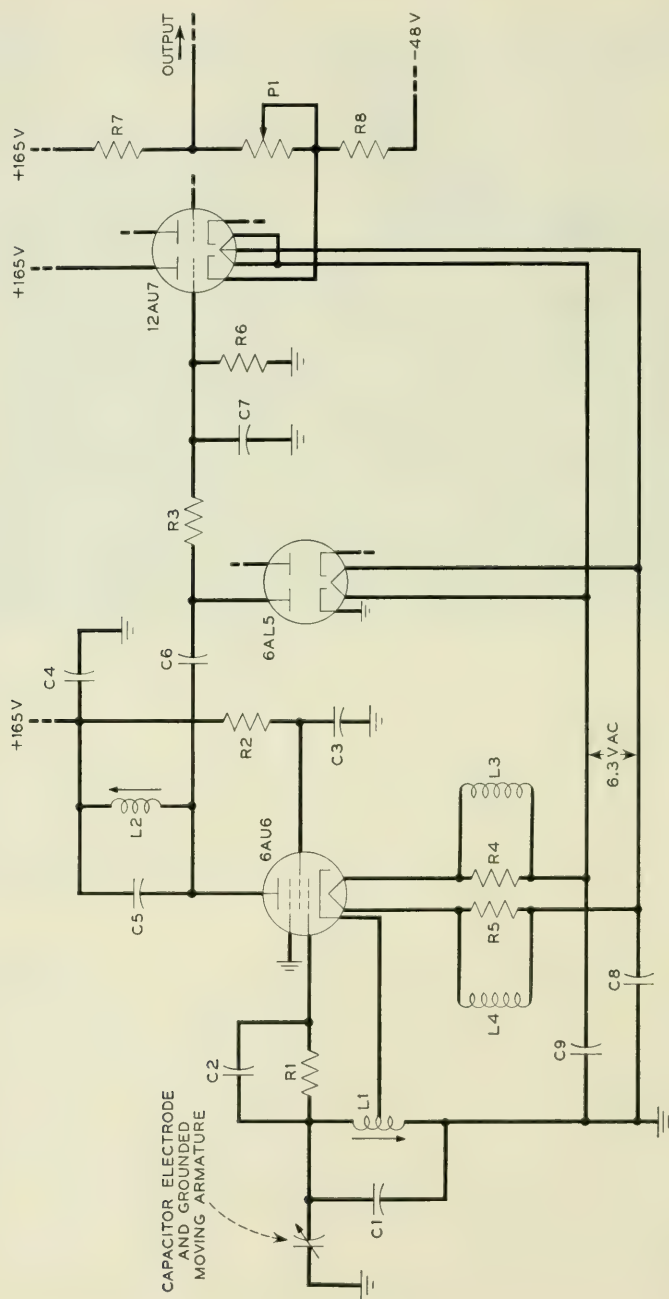


Fig. 3 — Schematic circuit of electrostatic gauge.

to 10 volts with 10 volts appearing when the armature is operated. A typical voltage-displacement curve is shown on Fig. 4. This output is applied to the grid of the "zero set" cathode follower and its bias is set so the output will be from approximately -2 volts to $+2$ volts. This voltage is applied to the horizontal dc amplifier of the scope.

Theory of Operation of Gauge

Since the capacitance probe is in the tank circuit the frequency of the oscillator changes with variations in the separation of the probe electrodes. Hence a displacement of the moving electrode causes a change in the frequency but no change in the amplitude of the alternating current through the pentode plate. The radio frequency voltage that appears across the coupling impedance is the product of the plate current and the coupling impedance.

Since the current is constant in amplitude:

$$E = I_0 Z,$$

where $E = rf$ voltage applied to rectifier, $I_0 = rf$ plate current from oscillator, and $Z =$ coupling impedance.

Since a change in E is due to a change in the oscillator frequency and the resultant change in Z ;

$$\text{Let } f = \frac{\omega}{2\pi} = \text{oscillator frequency,}$$

$$f_2 = \frac{\omega_2}{2\pi} = \text{resonance frequency of coupling circuit,}$$

$$Q = \frac{\omega_2 L_2}{R},$$

and

$$Y = 1 - \frac{f^2}{f_2^2}.$$

where $L_2 =$ inductance of coupling circuit, and $R =$ resistance of coupling circuit.

The coupling impedance can be written:

$$Z = \frac{\omega L_2 Q}{\sqrt{Q^2 Y^2 + 1}} \quad (1)$$

$$\text{At resonance } Z_{\max.} = \omega_2 L_2 Q.$$

Since the maximum value of E is obtained when Z is at a maximum:

$$\frac{E}{E_{\max}} = \frac{Z}{Z_{\max}} = \frac{1}{\sqrt{Q^2 Y^2 + 1}}, \quad (2)$$

provided ω/ω_2 is near unity.

Probe and Grid Circuit of Oscillator

The oscillator frequency is determined by the resonant frequency of the tank circuit. Fig. 3 shows that the tank comprises the capacitance of the probe plus a fixed capacitance and an inductance. The probe which is assumed to be a parallel plate condenser has a capacitance which is inversely proportional to the plate separation.

Let: d be plate separation of probe, d_1 be separation at resonance (where $f = f_2$),

$$x = \frac{d}{d_1}.$$

Then the variable capacitance of probe is given by:

$$C_p = C_1 \frac{d_1}{d} = \frac{C_1}{x}$$

and the total capacitance of the tank by:

$$C_0 + C_p = C_0 + \frac{C_1}{x},$$

where C_0 = fixed capacitance of tank including fixed part of probe capacitance.

Let:

$$\frac{C_1}{C_0} = K,$$

then

$$\omega^2 = \frac{1}{L_1 C_0 \left(1 + \frac{K}{x}\right)} \quad \text{and} \quad \omega_2^2 = \frac{1}{L_1 C_0 (1 + K)},$$

where ω and ω_2 are radian frequencies of the oscillator at probe separations d and d_1 respectively and L_1 is the effective inductance of the tank circuit. It follows that:

$$\frac{\omega^2}{\omega_2^2} = \frac{1 + K}{1 + \frac{K}{x}},$$

and therefore:

$$\begin{aligned} Y &= 1 - \frac{\omega^2}{\omega_2^2} = 1 - \frac{x + Kx}{K + x}, \\ &= \frac{1 - x}{1 + \frac{x}{K}}. \end{aligned} \quad (3)$$

This expression for Y may be substituted in (2) giving:

$$\frac{E}{E_{\max}} = \frac{1}{\sqrt{Q^2 Y^2 + 1}} = \frac{1}{\sqrt{1 + Q^2 \left(\frac{1 - x}{1 + \frac{x}{K}} \right)^2}}. \quad (4)$$

For $x/K \gg 1$ (4) may be written:

$$\frac{E}{E_{\max}} = \frac{1}{\sqrt{1 + K^2 Q^2 \left(\frac{1 - x}{x} \right)^2}}. \quad (5)$$

Equation (5) gives the voltage across the rectifier as a function of probe separation and the product KQ . It can be used to determine the optimum values of: (1) K the ratio of variable to fixed oscillator capacitance, (2) the Q of the coupling impedance, and (3) the range of probe

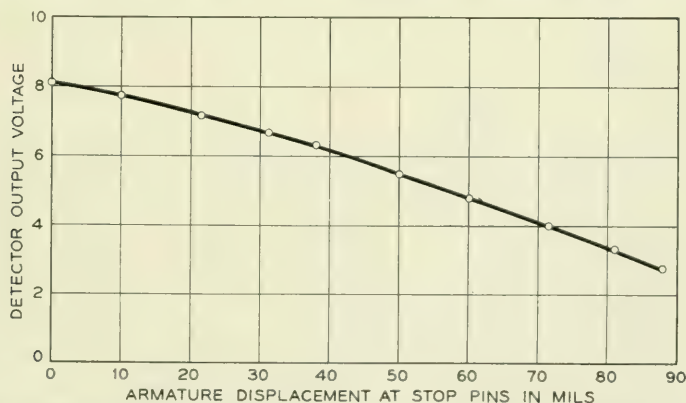


Fig. 4 — Typical output voltage versus displacement curve of the electrostatic gauge.

separations over which a prescribed degree of linearity can be obtained. It can be shown that the maximum range of linearity is obtained with $(KQ)^2 = 2$ and that the center of this range is at $x = \frac{1}{2}$. Fig. 5 is a graph of (5) against x with $(KQ)^2$ set at 2.

SCANNING CIRCUIT

The scanning circuit is a high speed electronic switching and detecting device which rapidly selects successive contacts, checks whether the contact is opened or closed, and displays this information on the vertical plates of the scope. Each contact is assigned a particular vertical reference level. When gauging, each level appears as a horizontal line with a vertical step on the line indicating the armature position for the contact operation.

Electron tubes designated "C" on Fig. 6 serve as switches to connect one contact at a time to the vertical plates of the scope. Each contact is connected to the grid circuit of a C tube. In order to test a contact its

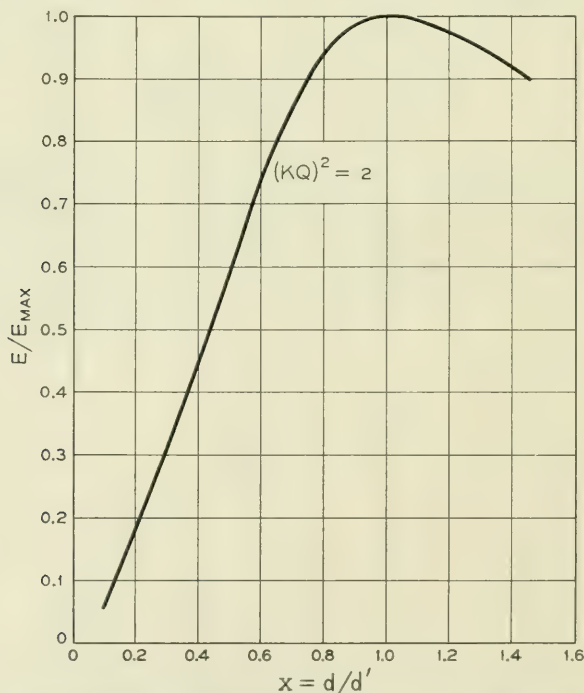


Fig. 5 — Curve showing E/E_{max} versus x with $(KQ)^2 = 2$ computed from equation (5).

associated C tube is made to conduct by shifting the grid voltage. All of the C tube cathodes are connected together to a common cathode resistor which couples the voltages from the contacts to the vertical amplifier of the scope. The plate current that flows when a C tube is fired causes a voltage drop through the common cathode resistor which deflects the scope beam to the proper vertical level. The level is set to the desired value by adjusting the plate circuit resistance. Since the contact under test is connected to the grid circuit, the grid voltage and the cathode voltage are shifted a small amount by opening or closing the contact. That is, when a C tube is fired the vertical plate voltage jumps to one value when the contact is open and it jumps to another slightly different value when the contact is closed. The 16 C tubes corresponding to the 16 contacts under test are fired one at a time in succession by 16 associated multivibrator stages. The multivibrators are connected in a ring so that each stage is fired by the preceding stage. When a stage is fired it holds its C tube in a conducting condition for two microseconds.

Fig. 6 is a detailed schematic of the single stage multivibrator (tubes A and B) with an associated modulator (C) tube. The multivibrators are normally in a stable waiting state and go into a temporary unstable state only when a transient is applied. Normally the A tube is cut off and the B tube is conducting. When a pulse is applied to the A tube grid the A tube conducts and the B tube is cut off. After two microseconds the A tube reverts back to its waiting state and sends a pulse to the next stage. When the B tube is cut off it provides a flat two microsecond pulse through the coupling circuit to the associated C tube.

Each multivibrator stage consisting of the A and B tubes with other circuit elements is mounted on a plug-in turret which may easily be changed in case of trouble.

BRIGHTNESS CONTROL CIRCUIT

When pulsing a relay the armature remains on the operated and released positions for a relatively long time interval. If the intensity of the scope beam is allowed to remain constant the spots at the ends of the traces are bright enough to fog the entire scope face. This makes it difficult to see the relatively weak lines that are caused by the motion of the armature. Therefore a control circuit is provided to brighten the scope trace only when the relay armature is in motion. The circuit shown on Fig. 8 provides the voltage required to intensify the cathode beam of the scope. This voltage is obtained by differentiating the output voltage of the electrostatic gauge. As the latter is proportional to the

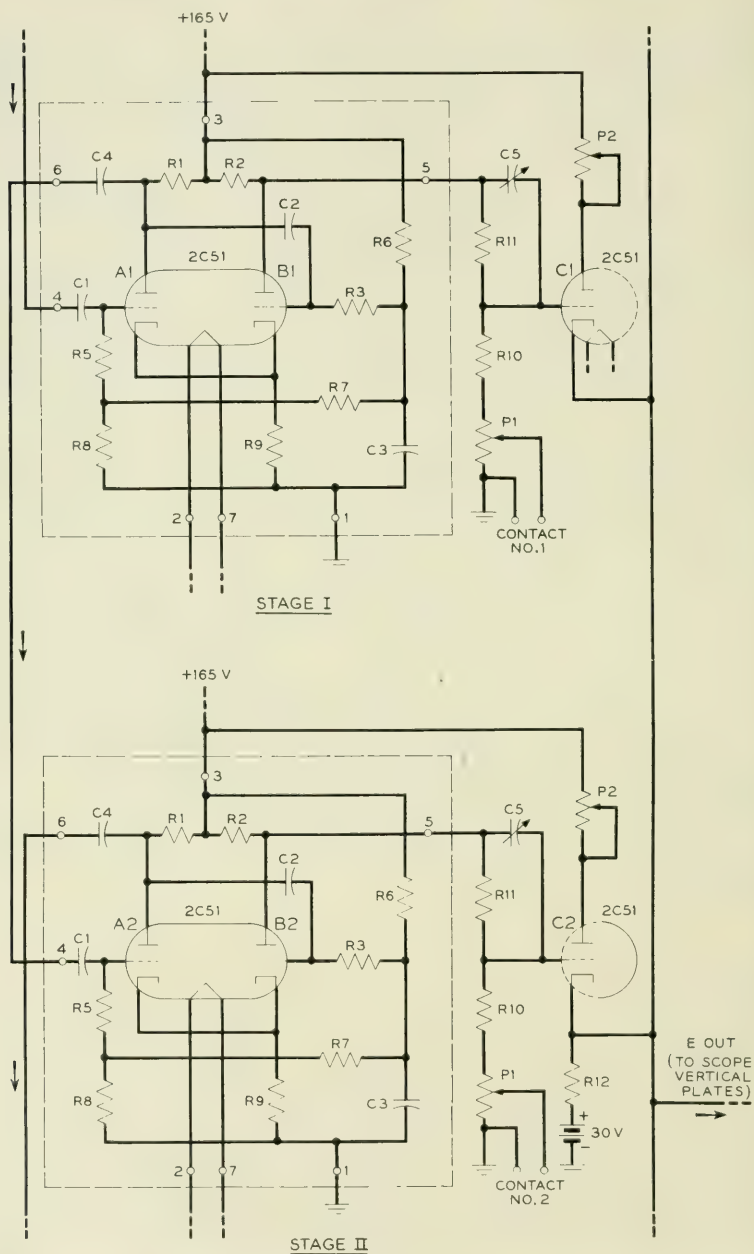


Fig. 6 -- Schematic circuit of two switching stages showing modulator tubes (C) and their associated multivibrator tubes (A and B).

armature position, the differentiated voltage is proportional to the armature velocity. After amplification the voltage from the gauge is applied to a tube with a transformer in the plate circuit. The voltage across the secondary coil of the transformer is the differentiated voltage, proportional to the armature velocity. It is applied to a germanium diode bridge, which may be connected as a half-wave or full-wave rectifier by the selector switch. The rectified voltage is amplified and clipped and then applied to the Z axis terminals of the scope. The scope trace is brightened during operate, during release, or during both in accordance with the setting of the bridge selector switch.

GAUGING A RELAY

The relay is plugged into a holding fixture with a jack for the coil and contact terminals. The jack connects the relay coil to the circuit which provides power for operating the relay. It also connects the relay contact terminals to the scanning circuit.

The electrostatic transducer electrode is mounted on a bracket which is attached to the front end of the fixture by means of a hinged bracket.

After the relay is plugged into the jack, it is clamped and the gauge electrode is rotated into position near the armature. Then the "zero set" potentiometer on the dc amplifier is used to align the operated armature position with the zero marking on the calibrated horizontal scale of the scope.

The contact selector switch is used to select the combination of contacts to be scanned such as: all contacts, all breaks, or all makes. For convenience in checking relays with 8 contacts or less a switch on the scanning circuit is used to connect 8 of the multivibrators in a ring instead of 16 so that only 8 lines appear on the scope. These 8 lines may be shifted to obtain greater spacing for easier reading. A typical gauging pattern on an oscilloscope screen is shown on Fig. 7.

This also improves the gauging accuracy which depends upon the amount of armature motion between successive scanning dots. For 16 horizontal lines on the screen the time interval between successive dots is 32 microseconds. For an armature velocity of 30 inches per second the corresponding distance between successive dots on one line is about 1 mil-inch. If 8 lines are used this distance is about 0.5 mil-inches. The gauging error resulting from this dot definition is actually less than indicated because successive operations of the armature give a small random variation in the dot locations.



Fig. 7 — Photograph of the 7-inch oscilloscope with a typical gauging pattern on the screen.

The brightness control switch is used to compare gauging when the relay is operating with the corresponding gauging when it is releasing. An apparent difference of 2 to 3 mil-inches is noticed between the operating gauge point of a contact and the releasing gauge point. This measurement change is ascribed to the tip flexure or follow of the contact wire. Contact closure occurs when the contacts first touch, in advance of the short follow travel during which the tension of the contact wire is transferred from the card to the contact. In dynamic operation, the contact opens when it is first struck by the card, before the tension would be transferred statically from the contact to the card. Thus the dynamic contact closure points agree with the static gauging, and differ from the dynamic opening point by the amount of contact follow. Good

correlation is obtained between the static and dynamic gauge measurements if make contacts are gauged on operate and break contacts on release.

OTHER RELAY CHARACTERISTICS

Switches are provided to permit other relay characteristics to be observed on the scope such as armature motion, contact chatter, and the contact operate and release times. The contact gauging is obtained with the horizontal plates connected to the electrostatic gauge. With this setting, the time for a dot to move across the screen may be less than 3 milliseconds (depending on the armature transit time) and contact chatter may be observed during this time.

With a time base sweep on the horizontal plates, the armature motion

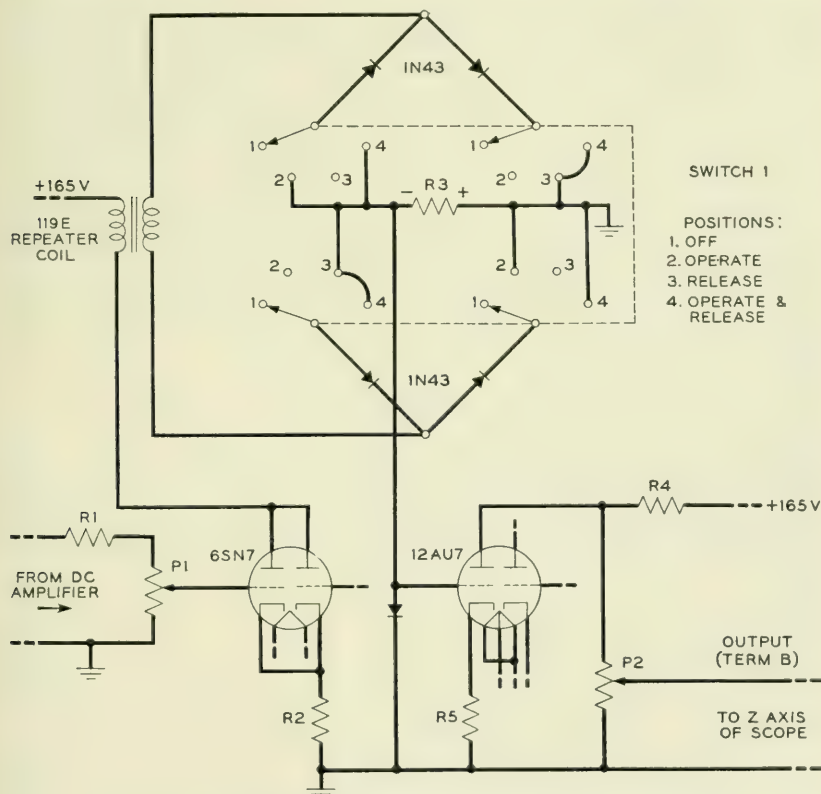


Fig. 8 — Schematic of oscilloscope brightening circuit.

may be put on the vertical plates. The chatter of 16 contacts may also be observed at one time to obtain a quick check of the contact performance. Operate and release times of the contacts may be read directly from the scope if a pip is put on the vertical axis when the test relay coil is energized. The front end of the relay is clamped for all measurements.

Switches are provided to select different combinations of contacts to be checked with either a time base horizontal sweep or with the armature displacement as the horizontal sweep.

DISCUSSION

The electronic relay tester was designed primarily for rapid inspection of wire spring relays. It provides a means for observing the position of the armature for the operation of all contacts simultaneously while the relay is pulsing. This method has several advantages over thickness gauges which are frequently used: (1) The contacts are gauged under conditions of use, that is, the armature moves as in normal operation without touching the gauging device; (2) Inspection is more rapid since the go-no-go positions are displayed on the scope face. Relay spring combinations requiring a sequence of contact operation are readily observed while the relay operates in a normal way; and (3) When relay adjustments are made the results of the adjustment are shown immediately so the contact operate points are centered within the desired range giving more margin for changes that may occur with use. This device is readily changed to measure other relay characteristics that may be of interest such as operate or release time, contact chatter and armature rebound.

ACKNOWLEDGMENTS

The authors are indebted to J. H. McConnell who designed a similar electrostatic transducer and to R. L. Peek, Jr., who derived the equations for it. Mr. Peek also suggested the method used in the scanning circuit.

Topics in Guided Wave Propagation Through Gyromagnetic Media

Part II—Transverse Magnetization and the Non-Reciprocal Helix

By H. SUHL and L. R. WALKER

(Manuscript received March 30, 1954)

Propagation through a gyromagnetic medium in a direction normal to a uniform magnetizing field is considered. Geometrical arrangements which make this propagation non-reciprocal are described. A few illustrative examples are discussed briefly. The non-reciprocal helix, of importance in traveling wave tube work, is treated at length.

1. INTRODUCTION

1.1. General Remarks about Non-Reciprocal Propagation

Part I of this paper began with a brief discussion of some of the microwave properties of two gyromagnetic media; the gas discharge plasma and the ferrite. The remainder of Part I was devoted to the analysis of the mode spectrum in a cylindrical waveguide filled with one of the media and placed in an axial magnetic field. It was demonstrated that the natural modes in such a guide are right- and left-circularly polarized waves which travel with different phase velocities. Accordingly a plane polarized mode, which to some approximation can be regarded as the sum of right and left circular modes, will, in traversing a section of the guide, undergo Faraday rotation, just like a plane wave in the unbounded medium. It is true that the presence of the guide wall has a drastic effect on the course of the rotation with magnetizing field, changing it, sometimes beyond recognition, from that prevailing in the unbounded medium. Nevertheless the principle remains the same; confinement of the wave to a guide merely modifies quantitatively the Faraday effect for plane waves. In optics, where practically plane waves are almost always employed in this connection, the non-reciprocal nature of this effect is so familiar that it hardly requires restatement here.

By contrast, Part II of this paper deals with devices whose non-reciprocal operation depends *in principle* as well as in numerical detail on the disposition of the boundary, or, more generally, on geometrical configuration. These devices employ magnetizing fields transverse to the propagation direction. Some electromagnetic field configurations are unaffected by such a dc field, but, whenever the rf magnetic field in the case of a ferrite (or the rf electric field for a plasma) has a component normal to the dc magnetic field, this is no longer the case. For, now, the magnetization (or the charges) will be caused to precess about the dc field, giving extra terms in Maxwell's equations and a resultant change in the propagation. This change may be simply an alteration in phase velocity, the propagation remaining reciprocal. This is the case, for example, for the propagation of plane waves in an infinitely extended medium [Cotton-Mouton effect]. Here, since every direction of propagation normal to the dc field is physically equivalent to any other and, in particular, to the opposite direction, no non-reciprocity can arise.

For reciprocity to be preserved in the presence of the dc magnetic field is, however, exceptional and requires a certain amount of geometrical symmetry in the system. That non-reciprocity may be expected in asymmetrical systems may be foreseen if we consider a system, typical of those to be treated in this paper, in which all the rf fields are independent of the coordinate along which the dc magnetic field is pointing. The relevant conducting boundaries and any interfaces between ferrite (plasma) and air are all surfaces parallel to the direction of propagation and lying in the dc magnetic field direction. Suppose the system to be divided into two parts by another surface of a similar kind and examine the surface impedance of one of the parts (which should contain some gyromagnetic material). If the propagation direction be reversed it is necessary to reverse the magnetic field to retrieve a situation in the part considered *geometrically* equivalent to the original. But, since the precession of the magnetization (or charges) about the dc magnetic field has a definite sense, the magnetic or electric current associated with this precession will be reversed when the dc field is reversed. Thus, the properties of the medium are altered and the surface impedance will be different for the two directions of propagation. In general, the surface impedance of the other part of the system will not compensate for this distinction between the two directions and we shall find different propagation constants for opposing directions. An exception will occur if the system contains a surface about which it has geometrical symmetry, for, then, compensation clearly takes place about this surface and the system is reciprocal.

An example of a simple non-reciprocal system is indicated in Fig. 1(a). Here a slab of ferrite is inserted into a rectangular waveguide parallel to the narrow walls and closer to one of them. Several workers have demonstrated that this arrangement and a similar arrangement in a circular waveguide are non-reciprocal for what is essentially the dominant mode.^{1, 3, *} When the slab is centered in the guide we have a plane of symmetry and the non-reciprocity vanishes.

Another configuration of the transverse field type is represented by the system shown in Fig. 1(b). Here a hollow ferrite cylinder is magnetized circumferentially and propagates a TE_{0n} -mode. It is clear that any arrangement of this sort, which might, in principle, include conducting sheaths, internally or externally, or might have the ferrite extending to an indefinitely large or small radius, cannot have any symmetry

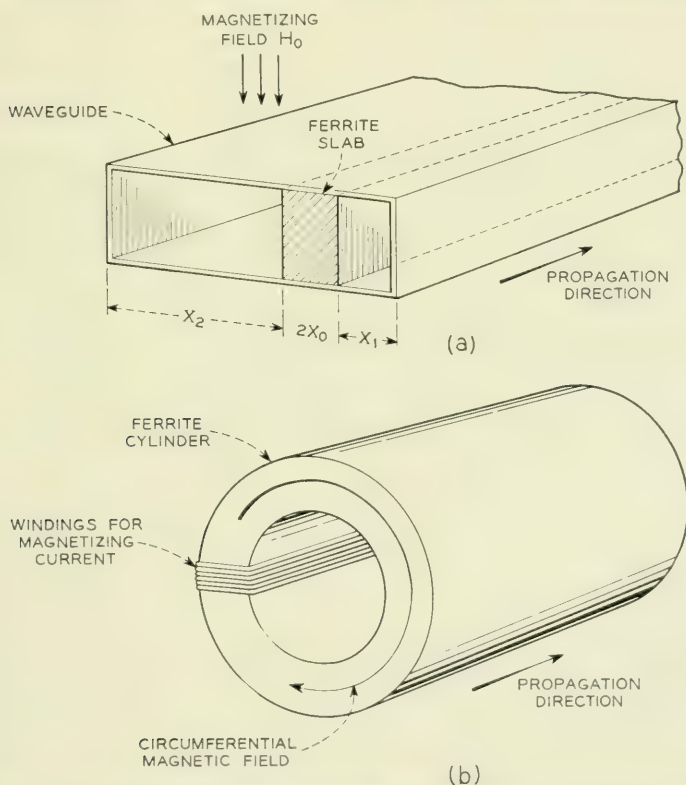


Fig. 1 -- (a) Rectangular waveguide and ferrite slab. (b) Circumferentially magnetized ferrite cylinder.

* It is expected that an article on this subject by S. E. Miller, A. G. Fox and M. T. Weiss will appear in a forthcoming issue of the JOURNAL.

about a cylinder coaxial with the ferrite. Thus, the internal and external impedances of such a system at any coaxial cylinder can only compensate accidentally (perhaps at a single frequency) and non-reciprocity is the rule.

It is frequently asserted without qualification that for non-reciprocity a further condition upon the relevant rf field is that its projection upon a plane normal to the magnetizing field be elliptically or circularly polarized in the limit of vanishing magnetization. The argument is based on the consideration, in itself correct, that the effective material constants are different for right- and left-circularly polarized field vectors. Suppose that the magnetization direction is y . Then the tensor relating B to H is (see Part I, Section 2.1):

$$\begin{vmatrix} \mu & 0 & -j\kappa \\ 0 & \mu_0 & 0 \\ j\kappa & 0 & \mu \end{vmatrix}.$$

For right- and left-circular fields with $H_z = \pm jH_x$, therefore, the medium is isotropic in the plane transverse to the dc field with permeabilities $\mu + \kappa$, $\mu - \kappa$ respectively. Since opposite circular polarizations accompany opposite propagation directions, (see for example, Fig. 2) the permeabilities, and hence the propagation constants, are different for opposite propagation directions. It is then argued that the field must already be circularly, or at least elliptically polarized to start with, if non-reciprocal effects are to result from application of the magnetization. However, the argument is true only for effects of first order in the magnetization. For general values of magnetization, the rf field, even if linearly polarized to begin with, will become elliptically polarized, and

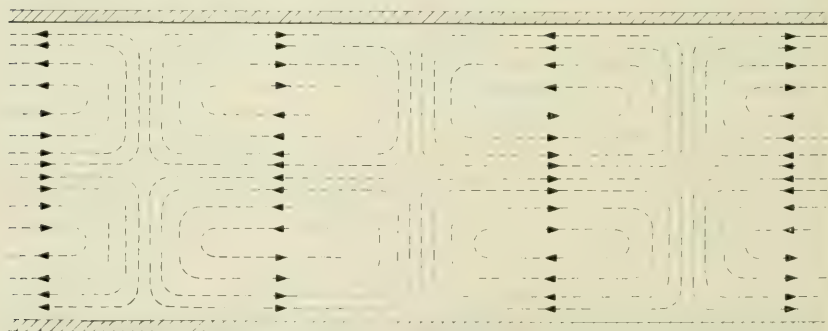


Fig. 2 — Magnetic lines of force parallel to the broad side of a rectangular waveguide.

non-reciprocity will occur. It is understandable that this double function of the magnetization (conversion to elliptic polarization plus creation of differences in permeability) leads to higher order effects. However, inasmuch as for the ferrite and plasma the magnetization can produce large changes, the requirement of elliptic polarization in zero magnetic field cannot be regarded as essential in practice. These considerations are demonstrated by a simple example in Section 2.2.

The devices considered in this paper actually are such that the electric or the magnetic field vector in the plane normal to the field is elliptically polarized even in the absence of the gyromagnetic medium. For example, the magnetic lines of force of the TE_{10} mode in a rectangular waveguide form two sets of closed loops in a plane parallel to the wide sides (see Fig. 2) and repeating every wavelength. This pattern moves bodily down the guide with the phase velocity of the mode, so that an observer stationary at any point not at the center or at the narrow walls of the guide sees a magnetic field rotating at the signal frequency and tracing out a generally elliptic path. The sense of the rotation is opposite on opposite sides of the center plane, and depends on the propagation direction. The conditions outlined in the previous paragraph are therefore satisfied; introduction of a ferrite slab magnetized as in Fig. 1(a) will yield first order non-reciprocal effects.

The problems considered here are such that the electromagnetic fields do not vary in the direction of magnetization. Under these conditions the field can be split into a TE and TM field satisfying different wave equations. In general, the two fields are coupled through the boundary conditions. Most of the paper is devoted to the analysis of the non-reciprocal helix, a problem that has recently gained importance in connection with high power traveling-wave-amplifiers.⁴ The conventional amplifier suffers from a limitation on its maximum useful gain; waves reflected from the output end will make the tube "sing" above a certain critical gain. Ferrites offer the possibility of preferentially suppressing these backward waves and so of increasing the permissible gain by a large amount.* In section 2.3 the "flat" helix (one of infinite radius) is considered. For the slow waves employed in practice a rather complete treatment is possible in this case of planar geometry. In Section 3 the cylindrical helix is treated. Inasmuch as the solutions involve functions for which no extensive tables exist, the treatment has to be more sketchy.

* More specifically, in high-power traveling-wave tubes the large beam current employed may be above the critical value required for backward wave oscillations due to spatial harmonics of the helix structure. In such cases the larger attenuation of backward waves will permit a higher beam current and therefore stable amplification to higher power levels.

Thus the loss is neglected and only the non-reciprocal phase characteristic is considered. The losses have then to be determined approximately by differentiation of the phase characteristics.

A few further problems were considered as illustrations of the general principles. One case, that of the plasma filling the space above an impedance sheet can actually be solved analytically and provides a particularly clear demonstration of the non-reciprocity. The case of the rectangular waveguide with a ferrite slab has already been considered extensively elsewhere, and only the results for a thin slab are given here. A problem with cylindrical symmetry is taken up in Section 3.3: a cylindrical waveguide fitted with a circumferentially magnetized cylinder

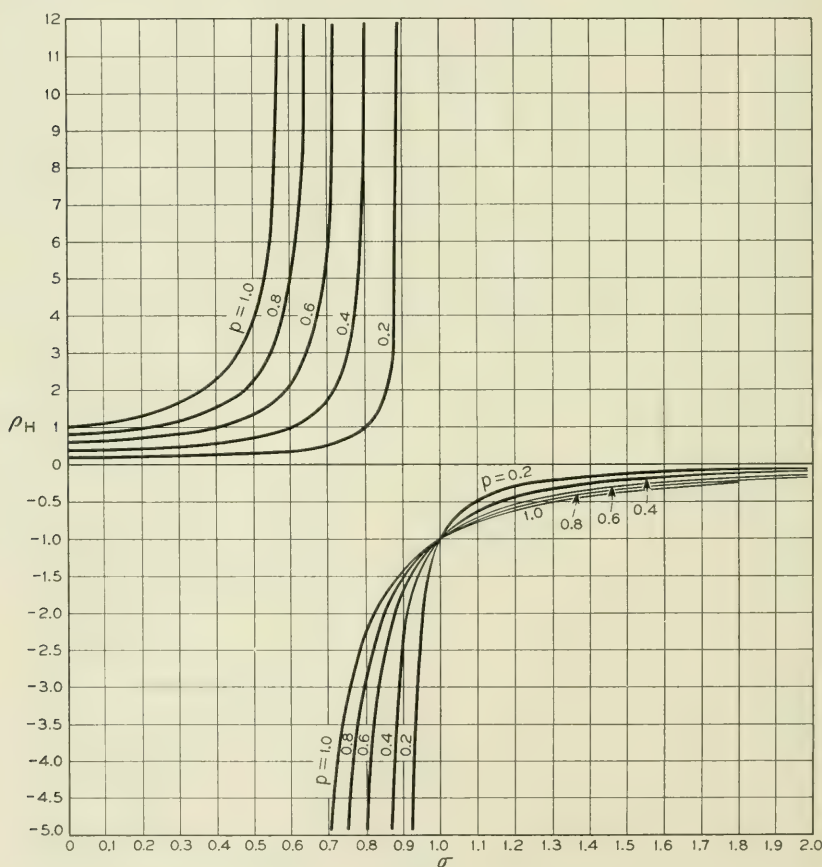


Fig. 3 — ρ_H versus σ for various p .

of ferrite. Again the discussion had to be sketchy in view of the scarcity of information on the functions that solve the problem.

2. PLANAR GEOMETRY

2.1. *Fields and Impedances*

In this section we consider planar transverse field problems which are characterized by the following conditions. The dc magnetic field is of uniform strength H_0 within the gyromagnetic medium and points along the y -axis. All rf field components are independent of the y coordinate. We discuss the ferrite case first, then indicate how the results are to be translated for the plasma.

For the orientation of the dc magnetic field which has been chosen the permeability matrix is of the form:

$$\begin{pmatrix} \mu & 0 & -j\kappa \\ 0 & \mu_0 & 0 \\ j\kappa & 0 & \mu \end{pmatrix},$$

μ and κ are, in general, even and odd functions of H_0 ; the permeability of unmagnetized ferrite is taken to be μ_0 as in free space. Following the procedure of Part I we shall assume specifically for μ and κ the formulae given by Polder's treatment of the dynamics of the medium. Thus, we have the expressions (for the case of no loss):

$$\begin{aligned} \frac{\mu}{\mu_0} &= \frac{1 - p\sigma - \sigma^2}{1 - \sigma^2}, \\ \frac{\kappa}{\mu_0} &= \frac{p}{1 - \sigma^2}, \text{ and} \\ \rho_H &= \frac{\kappa}{\mu} = \frac{p}{1 - p\sigma - \sigma^2}, \end{aligned} \tag{1}$$

where σ is the ratio of the precession frequency, $\frac{|\gamma|}{2\pi} H_0$, to the signal frequency and p is the ratio of a frequency, $\frac{|\gamma|}{2\pi} M_0/\mu_0$, associated with the saturation magnetization, M_0 , to the signal frequency. It should be noted that p and σ have always the same sign. The behavior of μ and κ as functions of σ was shown in Fig. 1(a) and (b) of Part I. ρ_H is shown as a function of σ in Fig. 3. The dielectric constant of the ferrite is taken to be ϵ . For reasons given in Part I, $|p|$ is assumed less than unity.

When the condition, $\partial/(\partial y) \equiv 0$, is put into Maxwell's equations the latter are found to be separable into two sets:

$$-\frac{\partial H_y}{\partial z} = j\omega\epsilon E_x, \quad (2a)$$

$$\frac{\partial H_y}{\partial x} = j\omega\epsilon E_z, \quad (2b)$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -j\omega\mu_0 H_y, \quad (2c)$$

and

$$-\frac{\partial E_y}{\partial z} = -j\omega[\mu H_x - j\kappa H_z], \quad (3a)$$

$$\frac{\partial E_y}{\partial x} = -j\omega[j\kappa H_x + \mu H_z], \quad (3b)$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = j\omega\epsilon E_y. \quad (3c)$$

It is to be stressed that such a separability is possible only when the rf fields do not vary along the dc magnetic field. The sets of equations (2) and (3) correspond to the separate equations for H_z and E_z which arise from (13) of Part I when β is there set equal to zero. The first set describes a TM field of the familiar type, whose propagation through the medium is unaffected by the magnetic field. The second set describes a TE field whose components, because of the presence of κ , are connected by different relations from those which exist in an unmagnetized medium. The separability of the two fields is equivalent to saying that they are not coupled by the medium itself, but they may, of course, be coupled at the boundaries.

We may write (3a) and (3b) in the form

$$-j\omega(\mu^2 - \kappa^2)H_x = j\kappa\frac{\partial E_y}{\partial x} - \mu\frac{\partial E_y}{\partial z}, \quad (4a)$$

$$-j\omega(\mu^2 - \kappa^2)H_z = \mu\frac{\partial E_y}{\partial x} + j\kappa\frac{\partial E_y}{\partial z}, \quad (4b)$$

and upon eliminating H_x and H_z , the wave equation for E_y is found to be:

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial z^2} + \omega^2\epsilon\frac{\mu^2 - \kappa^2}{\mu}E_y = 0, \quad (5)$$

where E_y (and also the H 's) are evidently propagated in the ferrite as

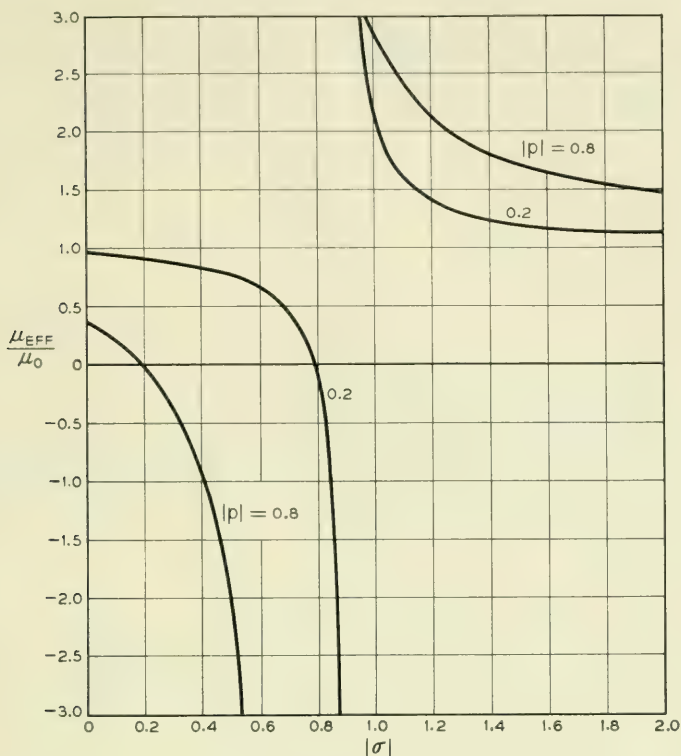


Fig. 4 — $\frac{\mu_{\text{eff}}}{\mu_0}$ versus $|\sigma|$.

though the latter had an effective permeability, $(\mu^2 - \kappa^2)/\mu = \mu_{\text{eff}}$. This permeability may assume any value from $+\infty$ to $-\infty$ as may be seen by writing it in terms of σ and p . From the Polder formulae, in fact,

$$\mu_{\text{eff}} = (\mu^2 - \kappa^2)/\mu = \mu_0 \frac{1 - (p + \sigma)^2}{1 - p\sigma - \sigma^2}. \quad (6)$$

As it should be, this is an even function of magnetic field. μ_{eff}/μ_0 decreases from $1 - p^2$ to 0 as $|\sigma|$ rises from 0 to $1 - |p|$; it decreases from 0 to $-\infty$ as $|\sigma|$ runs from $1 - |p|$ to $\sqrt{1 + p^2}/4 - |p|/2$ and finally decreases from ∞ to 1 as $|\sigma|$ increases indefinitely above $\sqrt{1 + p^2}/4 - |p|/2$. Its behavior is indicated in Fig. 4. It should be recalled from Part I that “ $\sigma = 0$ ” is an abbreviation for the very small magnetic field necessary to saturate the ferrite.

A brief examination of the propagation of plane waves shows the more

significant features of transmission in this medium. Since no direction in the x - z plane can be a preferred one, the plane wave may be assumed to travel in the z -direction with variation, $e^{-j\beta z}$. Then, from equations (4) and (5)

$$\beta^2 = \omega^2 \epsilon \mu_{\text{eff}},$$

and

$$H_x = \frac{-\beta \mu}{\omega(\mu^2 - \kappa^2)} E_y,$$

$$H_y = \frac{j\beta \kappa}{\omega(\mu^2 - \kappa^2)} E_y,$$

Since μ_{eff} is negative between $|\sigma| = 1 - |p|$ and $|\sigma| = \sqrt{1 + p^2/4} - |p|/2$, the medium is cut off for plane waves in this range of magnetic field (at a fixed signal frequency). H is elliptically polarized when the medium is magnetized and $|H_x|/|H_z| = |\kappa|/|\mu| = |\rho_H|$. We may also put

$$H_x + jH_z = \frac{-\beta E_y}{\omega(\mu - \kappa)},$$

$$H_x - jH_z = \frac{-\beta E_y}{\omega(\mu + \kappa)},$$

But $|H_x + jH_z|$ and $|H_x - jH_z|$ are proportional to the amplitudes of the left-handed and right-handed circularly polarized components of the magnetic field. The medium may thus be considered to exhibit the permeability,

$$\mu - \kappa = \mu_0 \left(1 - \frac{p}{1 - \sigma} \right),$$

for left-handed components and the permeability,

$$\mu + \kappa = \mu_0 \left(1 + \frac{p}{1 + \sigma} \right)$$

for right-handed components. The effective permeability is essentially a parallel combination of these two permeabilities and the medium may propagate ($\mu_{\text{eff}} > 0$) even when $\mu - \kappa$ is negative, as will happen for $\sqrt{1 + p^2/4} - |p|/2 < \sigma < 1$.

Since the medium itself has no non-reciprocal properties it is clear that if the latter are to arise they must do so as a result of interaction between the medium and its surroundings. The boundaries of the ferrite

at which matching will be necessary are surfaces parallel to the y -axis and here we need an expression for H_{tang}/E_y where H_{tang} is a tangential magnetic field at the surface. For the moment we will consider the admittance looking into the ferrite and take tangential components in the counterclockwise sense. From equation (4), then,

$$\frac{H_{\text{tang}}}{E_y} = \frac{j\mu \frac{1}{E_y} \frac{\partial E_y}{\partial \nu}}{\omega(\mu^2 - \kappa^2)} - \frac{\kappa \frac{1}{E_y} \frac{\partial E_y}{\partial \sigma}}{\omega(\mu^2 - \kappa^2)}, \quad (7)$$

where $\partial/(\partial \nu)$ is a normal derivative (outward) and $\partial/(\partial \sigma)$, a tangential derivative. It is possible, although by no means essential, to interpret the terms of (7) in the following fashion: the first term is just the admittance of a normal TE mode propagating in the interior of the ferrite (which is to have the permeability, μ_{eff}); the second term is to be ascribed to an independent surface current,

$$-\kappa \frac{\partial E_y}{\partial \sigma}.$$

Using this picture one may see how non-reciprocity arises in a simple case. If the ferrite be bounded by the planes $x = x_1$ and $x = x_2$, and the z -variation is of the form $e^{-j\beta z}$, the admittances due to the surface currents are $+j\kappa\beta/\omega(\mu^2 - \kappa^2)$. If β reverses its sign the surface currents are interchanged. If now the external admittances on the two sides of the slab are unequal (and, of course, themselves reciprocal) for given values of ω and $|\beta|$, there is no reason why β and $-\beta$ should simultaneously solve the matching problem.

Almost all the above considerations may be taken over to the case of the plasma. Here the TE fields will be undisturbed by the magnetic field (but the dielectric constant is altered by the presence of the charge from its free space value). Equations (4) and (5) are now replaced by

$$j\omega(\epsilon^2 - \eta^2)E_x = -\epsilon \frac{\partial H_y}{\partial z} + j\eta \frac{\partial H_y}{\partial x}, \quad (8a)$$

$$j\omega(\epsilon^2 - \eta^2)E_z = j\eta \frac{\partial H_y}{\partial z} + \epsilon \frac{\partial H_y}{\partial x}, \quad (8b)$$

$$\frac{\partial^2 H_y}{\partial x^2} + \frac{\partial^2 H_y}{\partial z^2} + \omega^2 \mu_0 \frac{\epsilon^2 - \eta^2}{\epsilon} H_y = 0 \quad (9)$$

for the TM fields. Here ϵ and η are the diagonal and off-diagonal terms of the dielectric matrix which is of the same form

$$\begin{vmatrix} \epsilon & 0 & -j\eta \\ 0 & \epsilon_1 & 0 \\ j\eta & 0 & \epsilon \end{vmatrix}$$

as the permeability matrix of the ferrite. The equations of motion for the plasma* lead to the expressions

$$\begin{aligned} \epsilon_1 &= \epsilon_0 (1 - q^2), \\ \epsilon &= \epsilon_0 \left(1 + \frac{q^2}{\sigma^2 - 1} \right), \\ \eta &= \epsilon_0 \frac{q^2 \sigma}{\sigma^2 - 1}, \\ \epsilon_{\text{eff}} &= \frac{\epsilon^2 - \eta^2}{\epsilon} = \epsilon_0 \frac{(1 - q^2)^2 - \sigma^2}{(1 - q^2) - \sigma^2}, \end{aligned} \quad (10)$$

where σ is now the ratio of the cyclotron resonance frequency;

$$\frac{1}{2\pi} \left| \frac{e}{m} \right| \mu_0 H_0,$$

in a field H_0 , to the applied frequency and q is the ratio of the plasma frequency to applied frequency; ϵ_{eff} behaves with magnetic field in much the same way as μ_{eff} . It is a constantly decreasing function of σ and is negative between $|\sigma| = 1 - q^2$ and $|\sigma| = \sqrt{1 - q^2}$, going to infinity at the latter value. The left and right handed dielectric constants, $\epsilon - \eta$ and $\epsilon + \eta$ are given by $\epsilon_0[1 - q^2/(1 - \sigma)]$ and $\epsilon_0[1 - q^2/(1 + \sigma)]$.

2.2. Examples of Non-reciprocal Systems

We now discuss briefly three examples of non-reciprocal systems as illustrations of particular points. As an example of a system which can be analyzed very easily and completely we consider a plasma occupying the region, $x > 0$ and bounded at $x = 0$ by a sheet of constant impedance. This impedance is to depend upon frequency but not on the propagation constant. It will be written as $j\sqrt{\mu_0/\epsilon_0}Z(\omega)$ where μ_0 and ϵ_0 are free space values. A practical realization of such a sheet might consist of a very large number of similar fins of negligible thickness and separation, parallel to the y -axis and attached normally to a conducting plane $x = \text{constant}$. The fields between separate fins are uncoupled and E_z is uniform between fins. For such an arrangement, $Z(\omega) = \tan \omega \sqrt{\epsilon_0 \mu_0} x_0$,

* See Section 2.2 of Part I.

where x_0 is the depth of the fins. If the z -variation of the fields is $e^{-j\beta z}$ and the waves are guided, the x -dependence in the plasma must be as $\exp -\sqrt{\beta^2 - \omega^2 \mu_0 \epsilon_{\text{eff}}} x$. From equation (8b)

$$j\omega(\epsilon^2 - \eta^2)E_z = [\eta\beta - \epsilon\sqrt{\beta^2 - \omega^2 \mu_0 \epsilon_{\text{eff}}}]H_y.$$

Matching at $x = 0$ gives

$$\frac{\eta\beta - \epsilon\sqrt{\beta^2 - \omega^2 \mu_0 \epsilon_{\text{eff}}}}{j\omega(\epsilon^2 - \eta^2)} = j\sqrt{\frac{\mu_0}{\epsilon_0}} Z(\omega).$$

This yields

$$\left[\beta - \eta\omega\sqrt{\frac{\mu_0}{\epsilon_0}} Z(\omega) \right]^2 = \omega^2 \mu_0 \epsilon + \omega^2 \frac{\mu_0}{\epsilon_0} \epsilon^2 Z^2(\omega)$$

or

$$\omega\sqrt{\frac{\beta}{\mu_0 \epsilon_0}} = \frac{\eta}{\epsilon_0} Z(\omega) \pm \sqrt{\frac{\epsilon}{\epsilon_0} + \frac{\epsilon^2}{\epsilon_0^2} Z^2(\omega)}. \quad (11)$$

The non-reciprocity is clearly exhibited, since the two values of β are not equal and opposite. The solution (11) is valid only if $\beta^2 > \omega^2 \mu_0 \epsilon_{\text{eff}}$, corresponding to guided waves.

In the second example we assume that the region between conducting planes at $x = 0$ and $x = x_0$ is filled by a plasma. When no magnetic field is present E_z is supposed to vanish and E_x is uniform across the gap. The unperturbed field is then plane polarized (TEM). The magnetic field is now applied parallel to the y axis so that part of the gap between $x = 0$ and $x = x_1$, E_z in the magnetized plasma is now given by

$$E_z = E_0 \sin \sqrt{\omega^2 \mu_0 \epsilon_{\text{eff}} - \beta^2} x$$

since it must vanish at $x = 0$. The z variation is again $\exp -j\beta z$. H_y may be found from equation (8b) and is

$$H_y = \frac{j\omega E_0}{\omega^2 \mu_0 \epsilon - \beta^2} [\eta\beta \sin \sqrt{\omega^2 \mu_0 \epsilon_{\text{eff}} - \beta^2} x - \epsilon \sqrt{\omega^2 \mu_0 \epsilon_{\text{eff}} - \beta^2} \cos \sqrt{\omega^2 \mu_0 \epsilon_{\text{eff}} - \beta^2} x].$$

The admittance of the magnetized section at $x = x_1$ is, thus,

$$\frac{H_y}{E_z} = \frac{j\omega}{\omega^2 \mu_0 \epsilon - \beta^2} [\eta\beta - \epsilon \sqrt{\omega^2 \mu_0 \epsilon_{\text{eff}} - \beta^2} \cot \sqrt{\omega^2 \mu_0 \epsilon_{\text{eff}} - \beta^2} x_1],$$

and, analogously, that of the unmagnetized part is

$$\frac{H_y}{E_z} = \frac{j\omega}{\omega^2 \mu_0 \epsilon_1 - \beta^2} [-\epsilon_1 \sqrt{\omega^2 \mu_0 \epsilon_1 - \beta^2} \cot \sqrt{\omega^2 \mu_0 \epsilon_1 - \beta^2} (x_0 - x_1)],$$

where $\epsilon_1 = \epsilon_0(1 - q^2)$. Suppose now that the applied field is weak, so that $\omega^2\mu_0\epsilon \sim \omega^2\mu_0\epsilon_1 \sim \beta^2$. Then, the cotangents may be expanded and by equating admittances one obtains

$$\frac{j\omega}{\omega^2\mu_0\epsilon - \beta^2} \left[\eta\beta - \frac{\epsilon}{x_1} \right] = \frac{j\omega\epsilon_1}{(\omega^2\mu_0\epsilon_1 - \beta^2)(x_0 - x_1)},$$

or

$$\frac{\beta^2}{\omega^2\mu_0\epsilon_1} = 1 + \frac{\epsilon - \epsilon_1}{-\epsilon_1 + \left(\eta\beta - \frac{\epsilon}{x_1} \right) (x_0 - x_1)}. \quad (12)$$

Since $\epsilon - \epsilon_1$ is of the second order in σ this equation may be solved by substituting the unperturbed value of β , $\pm \omega\sqrt{\mu_0\epsilon_1}$, in the right hand side. It is clear that although the system is non-reciprocal, it will be so only to third order in σ . This system, therefore, illustrates the fact pointed out in Section 1.1 that even when the fields are plane polarized in the absence of a dc magnetic field, non-reciprocity may arise, although it may be very small in weak fields.

The third example to be considered is one which has been referred to in Section 1.1, namely that in which a strip of ferrite is placed across the short dimension of rectangular wave guide, see Fig. 1(a). In view of the fact that this problem has been discussed with great thoroughness by Lax, Button and Roth, we shall, after deriving the characteristic equation, consider only the case of a very thin strip. Let the thickness of the strip be $2x_0$ and the distance from its two faces to the nearest guide wall be x_1 and x_2 respectively. The admittances at the two faces are then

$$\frac{j}{\omega\mu_0} \cot \sqrt{\beta_0^2 - \beta^2} x_1$$

and

$$\frac{-j}{\omega\mu_0} \cot \sqrt{\beta_0^2 - \beta^2} x_2$$

respectively, where $\beta_0^2 = \omega^2\mu_0\epsilon_0$. Inside the ferrite

$$\frac{\partial^2}{\partial x^2} \equiv -[\omega^2\epsilon\mu_{\text{eff}} - \beta^2] = -(\beta_f^2 - \beta^2),$$

and

$$-j\omega(\mu^2 - \kappa^2)H_z = \mu \frac{\partial E_y}{\partial x} + \kappa\beta E_y.$$

Thus, immediately within the ferrite,

$$\frac{1}{E_y} \frac{\partial E_y}{\partial x} = -j\omega\mu_{\text{eff}} \left(\frac{H_z}{E_y} \right)_{\text{external}} - \rho_H \beta.$$

If the two faces of the ferrite are $x = -x_0$ and $x = x_0$ we then have

$$\left(\frac{1}{E_y} \frac{\partial E_y}{\partial x} \right)_{x=-x_0} = \frac{\mu_{\text{eff}}}{\mu_0} \cot \sqrt{\beta_0^2 - \beta^2} x_1 - \rho_H \beta = A,$$

and

$$\left(\frac{1}{E_y} \frac{\partial E_y}{\partial x} \right)_{x=x_0} = -\frac{\mu_{\text{eff}}}{\mu_0} \cot \sqrt{\beta_0^2 - \beta^2} x_2 - \rho_H \beta = B.$$

If we write

$$\frac{1}{E_y} \frac{\partial E_y}{\partial x} = \sqrt{\beta_f^2 - \beta^2} \tan (\sqrt{\beta_f^2 - \beta^2} x),$$

and make use of the boundary conditions we obtain

$$\tan 2 \sqrt{\beta_f^2 - \beta^2} x_0 = \frac{\sqrt{\beta_f^2 - \beta^2} (A - B)}{\beta_f^2 - \beta^2 - AB}. \quad (13)$$

The non-reciprocity is clearly contained in the odd power of β in A and B .

For small thickness we replace the tangent by its argument, and, substituting for A and B on one side of the equation, obtain

$$\frac{\mu_{\text{eff}}}{\mu_0} [\cot \sqrt{\beta_0^2 - \beta^2} x_1 + \cot \sqrt{\beta_0^2 - \beta^2} x_2] = 2x_0[\beta_f^2 - \beta^2 - AB],$$

or

$$\begin{aligned} & \sin \sqrt{\beta_0^2 - \beta^2} (x_1 + x_2) \\ &= 2x_0 \frac{\mu_0}{\mu_{\text{eff}}} \sin \sqrt{\beta_0^2 - \beta^2} x_1 \sin \sqrt{\beta_0^2 - \beta^2} x_2 [\beta_f^2 - \beta^2 - AB]. \end{aligned} \quad (14)$$

Since the guide is almost empty, we may write

$$\sqrt{\beta_0^2 - \beta^2} (x_1 + x_2) = \pi - \delta,$$

where δ is small. Or,

$$\beta = \beta_1 + \frac{2\pi\delta}{\beta_1(x_1 + x_2)^2},$$

where

$$\beta_1^2 = \beta_0^2 - \pi^2/(x_1 + x_2)^2.$$

Writing $\beta = \beta_1$ in the right hand side of equation (14) and noting that if

$$\sqrt{\beta_0^2 - \beta_1^2} x_1 = \frac{\pi x_1}{x_1 + x_2} = \theta$$

then $\sqrt{\beta_0^2 - \beta_1^2} x_2 = \pi - \theta$, we have

$$\begin{aligned} \delta &= 2x_0 \frac{\mu_0}{\mu_{\text{eff}}} \sin^2 \theta \left[\beta_f^2 - \beta_1^2 - \left(\frac{\mu_{\text{eff}}}{\mu_0} \cot \theta - \rho_H \beta_1 \right) \left(\frac{\mu_{\text{eff}}}{\mu_0} \cot \theta - \rho_H \beta_1 \right) \right], \\ &= 2x_0 \frac{\mu_0}{\mu_{\text{eff}}} \left[\left(\beta_f^2 - (1 + \rho_H^2) \beta_1^2 \right) \sin^2 \theta - \left(\frac{\mu_{\text{eff}}}{\mu_0} \right)^2 \cos^2 \theta \right. \\ &\quad \left. + 2 \frac{\mu_{\text{eff}}}{\mu_0} \rho_H \beta_1 \cos \theta \sin \theta \right]. \end{aligned} \quad (15)$$

The non-reciprocal part of β is thus $4\pi x_0(x_1 + x_2)^{-2} \rho_H \sin 2\theta$. This has a maximum value for $\theta = \pi/4$ or $3\pi/4$ and, hence, $x_1 = (x_1 + x_2)/4$ or $3(x_1 + x_2)/4$. This result may be understood qualitatively by considering the fields in the guide before the ferrite is inserted. We then have $E_y = E_0 \sin(\pi x/a)$ where $a = x_1 + x_2$ and consequently,

$$H_x = -\frac{\beta}{\omega\mu} \sin \frac{\pi x}{a} \quad \text{and} \quad H_z = j \frac{\pi}{a} \cos \frac{\pi x}{a}.$$

The amplitudes of the left- and right-handed components of circular polarization at a point are then proportional to

$$-\beta \sin \frac{\pi x}{a} - \frac{\pi}{a} \cos \frac{\pi x}{a} \quad \text{and} \quad -\beta \sin \frac{\pi x}{a} + \frac{\pi}{a} \cos \frac{\pi x}{a}.$$

The difference in the squares of these amplitudes is $2\beta(\pi/a) \sin(\pi x/a) \cos(\pi x/a)$ and this is proportional to the difference between the energy stored at x in the left-handed wave and in the right-handed wave. It is plausible that this should be a measure of the non-reciprocal effect produced by a thin piece of ferrite at x .

2.3. The Plane Helix

In dealing with transverse field problems with cylindrical geometry we shall consider non-reciprocal propagation along a helix which is surrounded by circumferentially magnetized ferrite. The analysis of this problem is rather cumbersome and it is advantageous to study first an analogous plane problem. The simplicity thus gained allows us to examine somewhat more complicated problems. The "plane helix,"

to be treated here, is a sheet of negligible thickness, lying in the plane $x = 0$, which conducts only in a single direction making an angle ψ with the y -axis. In this direction it will be supposed lossless. In addition we assume that the regions, $x < 0$ and $0 < x < x_0$ are empty, while the space $x > x_0$ is filled with ferrite. As usual the magnetic field is along the y -axis and the fields are independent of y . The problem is clearly the limit for very large radius of that of an empty, helically-conducting cylinder, surrounded by an infinitely thick shell of ferrite, circumferentially magnetized, the whole system carrying fields with no angular variation.

We first consider the boundary conditions for the plane helix, after noting that it is evident that both TE and TM fields will be required. The tangential electric field on either side of the sheet must necessarily be at right angles to the direction of conduction since the conductivity is infinite. Further, the tangential electric field must be continuous through the sheet. Hence, if the field normal to the direction of conduction is E_0 (omitting here and elsewhere the factor $e^{-j\beta z}$), we have

$$\begin{aligned} E_z^+ &= E_z^- = E_0 \cos \psi, \\ E_y^+ &= E_y^- = -E_0 \sin \psi, \end{aligned}$$

where the symbols $+$ and $-$ refer to $x > 0$ and $x < 0$ respectively. Again, since current cannot flow normal to the direction of conduction, the tangential magnetic field along the latter must be continuous through the sheet or

$$(H_z^+ - H_z^-) \sin \psi + (H_y^+ - H_y^-) \cos \psi = 0.$$

The boundary conditions may be combined into a single equation, by introducing admittances, in the form

$$\left(\frac{H_z^+}{E_y^+} - \frac{H_z^-}{E_y^-} \right) \sin^2 \psi = \left(\frac{H_y^+}{E_z^+} - \frac{H_y^-}{E_z^-} \right) \cos^2 \psi. \quad (16)$$

The left-hand side refers to the TE fields and the right to TM fields. In the empty regions surrounding the sheet

$$\left[\frac{\partial^2}{\partial x^2} - (\beta^2 - \omega^2 \epsilon_0 \mu_0) \right] H_y = \left[\frac{\partial^2}{\partial x^2} - (\beta^2 - \beta_0^2) \right] H_y = 0,$$

$$E_z = \frac{1}{j\omega\epsilon_0} \frac{\partial H_y}{\partial x},$$

and

$$\left[\frac{\partial^2}{\partial x^2} - (\beta^2 - \beta_0^2) \right] E_y = 0, \quad H_z = \frac{-1}{j\omega\mu_0} \frac{\partial E_y}{\partial x},$$

with $\beta_0^2 = \omega^2 \epsilon_0 \mu_0$. If the waves are to be guided, $\beta^2 > \beta_0^2$, and, then, for $x < 0$, we have $\partial/\partial x \equiv \sqrt{\beta^2 - \beta_0^2}$. Thus

$$\frac{H_y^-}{E_z^-} = \frac{j\omega\epsilon_0}{\sqrt{\beta^2 - \beta_0^2}},$$

and

$$\frac{H_z^-}{E_y^-} = - \frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\mu_0}.$$

If the admittances at the surface of the ferrite are H_y^f/E_z^f and H_z^f/E_y^f , then H_y^+/E_z^+ and H_z^+/E_y^+ are given by the impedance transformation:

$$\frac{H_y^+}{E_z^+} = \frac{j\omega\epsilon_0}{\sqrt{\beta^2 - \beta_0^2}} \cdot \frac{\frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\epsilon_0} \frac{H_y^f}{E_z^f} - \tanh \sqrt{\beta^2 - \beta_0^2} x_0}{1 - \frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\epsilon_0} \frac{H_y^f}{E_z^f} \cdot \tanh \sqrt{\beta^2 - \beta_0^2} x_0},$$

and

$$\frac{H_z^+}{E_y^+} = - \frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\epsilon_0} \cdot \frac{- \frac{j\omega\mu_0}{\sqrt{\beta^2 - \beta_0^2}} \frac{H_z^f}{E_y^f} - \tanh \sqrt{\beta^2 - \beta_0^2} x_0}{1 - \left[\frac{-j\omega\mu_0}{\sqrt{\beta^2 - \beta_0^2}} \frac{H_z^f}{E_y^f} \right] \cdot \tanh \sqrt{\beta^2 - \beta_0^2} x_0}.$$

Within the ferrite the TM-fields fall off as $\exp - \sqrt{\beta^2 - \omega^2 \epsilon_0 \mu_0} x$ and the TE-fields as $\exp - \sqrt{\beta^2 - \omega^2 \epsilon_0 \mu_{\text{eff}}} x$. We then have

$$\frac{H_y^f}{E_z^f} = \frac{-j\omega\epsilon}{\sqrt{\beta^2 - \omega^2 \epsilon \mu_0}},$$

and

$$\frac{H_z^f}{E_y^f} = \frac{j\omega\mu_{\text{eff}}}{\sqrt{\beta^2 - \omega^2 \epsilon \mu_{\text{eff}} - \beta \kappa / \mu}}.$$

These results may now be collected and substituted in equations (16). The equation of condition so obtained is

$$\begin{aligned} & (\beta^2 - \beta_0^2) \frac{A + 1}{A + \tanh \sqrt{\beta^2 - \beta_0^2} x_0} \\ &= \beta_0^2 \cot^2 \psi \frac{1 - B}{1 - B \tanh \sqrt{\beta^2 - \beta_0^2} x_0}, \end{aligned} \quad (17)$$

where

$$A = \frac{\mu_{\text{eff}}}{\mu_0} \frac{\sqrt{\beta^2 - \beta_0^2}}{\sqrt{\beta^2 - \omega^2 \epsilon \mu_{\text{eff}} - \beta \rho_H}},$$

and

$$B = -\frac{\epsilon}{\epsilon_0} \frac{\sqrt{\beta^2 - \beta_0^2}}{\sqrt{\beta^2 - \omega^2 \epsilon \mu_0}}.$$

We shall assume that we are dealing with slow waves (β large). This is the case of greatest practical interest and is usually ensured by making $\cot \psi$ large. Assuming that the waves are slow we simplify the equation (12) and find certain values for β . We may then ascertain in what ranges of σ and p the simplifying assumptions and the results are consistent.

In equation (17), then, we replace all square roots by $|\beta|$ and $\beta^2 - \beta_0^2$ by β^2 , obtaining

$$\beta^2 = \beta_0^2 \cot^2 \psi \frac{A + \tanh |\beta| x_0}{A + 1} \cdot \frac{1 - B}{1 - B \tanh |\beta| x_0},$$

with

$$A = \frac{\mu_{\text{eff}}/\mu_0}{1 - \rho_H \operatorname{sgn} \beta} = \frac{\mu + \kappa \operatorname{sgn} \beta}{\mu_0},$$

and

$$B = -\epsilon/\epsilon_0$$

where $\operatorname{sgn} \beta = 1$ for $\beta > 0$ and $\operatorname{sgn} \beta = -1$ for $\beta < 0$. Taking first the simplest case of no separation ($x_0 = 0$), this becomes

$$\beta^2 = \frac{\beta_0^2(1 + \epsilon/\epsilon_0) \cot^2 \psi}{1 + \frac{1 - \rho_H \operatorname{sgn} \beta}{\mu_{\text{eff}}/\mu_0}} = \frac{\beta_0^2(1 + \epsilon/\epsilon_0) \cot^2 \psi}{1 + \frac{\mu_0}{\mu + \kappa \operatorname{sgn} \beta}}.$$

Substituting the expressions (6) and (1) for μ_{eff}/μ_0 and ρ_H we are led to*

$$\beta_+ = \beta_0 \cot \psi \cdot \sqrt{\frac{1}{2}(1 + \epsilon/\epsilon_0)} \sqrt{\frac{\sigma + 1 + p}{\sigma + 1 + p/2}}, \quad (18a)$$

$$\beta_- = -\beta_0 \cot \psi \cdot \sqrt{\frac{1}{2}(1 + \epsilon/\epsilon_0)} \sqrt{\frac{\sigma - 1 + p}{\sigma - 1 + p/2}} \quad (18b)$$

* Since reversal of the magnetization is equivalent to interchange of the propagation directions, we are at liberty to consider σ and p always positive, and to deal with the two cases $\beta > 0$ and $\beta < 0$ separately.

The β_+ mode propagates for all σ , $\frac{\sigma + 1 + p}{\sigma + 1 + p/2}$ declining steadily from $\frac{1 + p}{1 + p/2}$ to unity. The β_- mode on the other hand is cut off, with $\beta^2 = 0$ at $\sigma = 1 - p$ and then reappears with $\beta^2 = \infty$ at $\sigma = 1 - p/2$. The behavior of the two modes is shown in Fig. 5. Self-consistency requires that $\beta^2 \gg \omega^2 \epsilon \mu_0$ or $\omega^2 \epsilon |\mu_{\text{eff}}|$, whichever is the greater. The condition $\beta^2 \gg \omega^2 \epsilon |\mu_{\text{eff}}|$ fails to be met near

$$\sigma = \sigma_0 = \sqrt{\frac{p^2}{4} + 1} - \frac{p}{2},$$

when

$$\frac{|\sigma - \sigma_0|}{\frac{p}{2} \sqrt{\frac{p^2}{4} + 1} \pm \left(\frac{p}{2} - 1\right)} < \frac{2}{(1 + \epsilon/\epsilon_0) \cot \psi}$$

The condition $\beta^2 \gg \omega^2 \epsilon \mu_0$ will fail for β_- near $\sigma = 1 - p$. The range for this to occur is given by

$$1 - p - \sigma < \frac{p/2}{\frac{1}{2} (1 + \epsilon/\epsilon_0) \cot \psi - 1}.$$

The extension of the Polder formulae to the case of a lossy ferrite was given in Part I, (Section 2.1). From the results given there one may write

$$\mu + \kappa \operatorname{sgn} \beta = 1 + p \frac{\sigma(1 + \alpha^2) - \operatorname{sgn} \beta + j\alpha}{\sigma^2(1 + \alpha^2) - 1 + 2j\alpha\sigma}$$

where α is a damping parameter. Substitution of these expressions in the slow wave formula for β^2 will give the effect of loss on the propagation constant. In Fig. 6 the complex value of β_- is shown for $\alpha = 0.1$ and several values of p . The imaginary part of β_+ is small and varies only slightly with σ . Fig. 7 shows the initial loss ($\sigma = 0$) for three values of α and a range of p values. From a knowledge of β as a function of $\sigma = \omega_H/\omega$ and $p = \omega_M/\omega$ it is possible to calculate the loss, $\operatorname{Im} \beta_- / (\beta_0 \sqrt{1 + \epsilon/\epsilon_0} \cot \psi)$, as a function of frequency when the magnetic field and saturation magnetization of the ferrite are held constant. Fig. 8 shows the results of such calculations. It should be noted that the horizontal scale is linear in σ or $1/\omega$ and that the vertical scale implicitly contains the frequency in the form of $1/\beta_0$. Both of these distortions of scale tend to

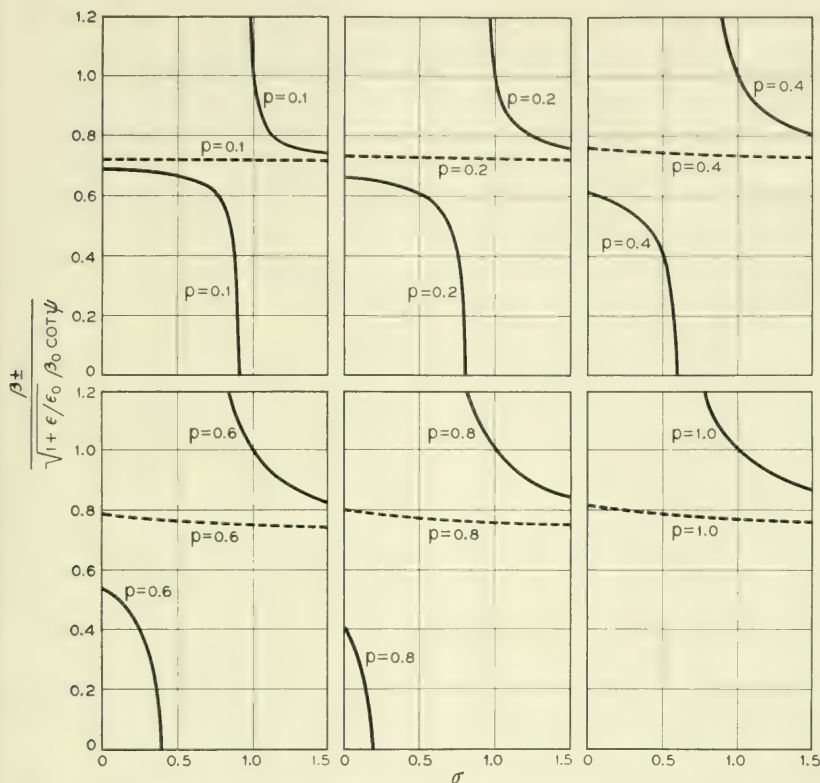


Fig. 5 — The non-reciprocal propagation constants of the flat helix. The dotted curves represent β_+ , the solid curves β_- . (Loss free case).

produce an appearance of sharpness in the variation of the loss at higher frequencies.

If the slow wave assumption be made again it is possible to obtain solutions for the case in which the ferrite does not touch the helix and the latter is lossless. With the slow wave approximation, (17) becomes

$$\frac{\beta_{\pm}^2}{\beta_0^2 \cot^2 \Psi} = \frac{1 + \frac{\epsilon}{\epsilon_0}}{1 + \frac{\epsilon}{\epsilon_0} \tanh |\beta| x_0} \cdot \frac{\frac{\mu + \kappa \operatorname{sgn} \beta}{\mu_0} + \tanh |\beta| x_0}{\frac{\mu + \kappa \operatorname{sgn} \beta}{\mu_0} + 1}$$

If we write $|\beta| x_0 = u$, then the above equation expresses $\beta_{\pm}$ in terms of u . At the same time $x_0 = u/|\beta_{\pm}(u)|$ and we evidently have a parametric representation of $\beta_{\pm}$ and x_0 . The results of such computations

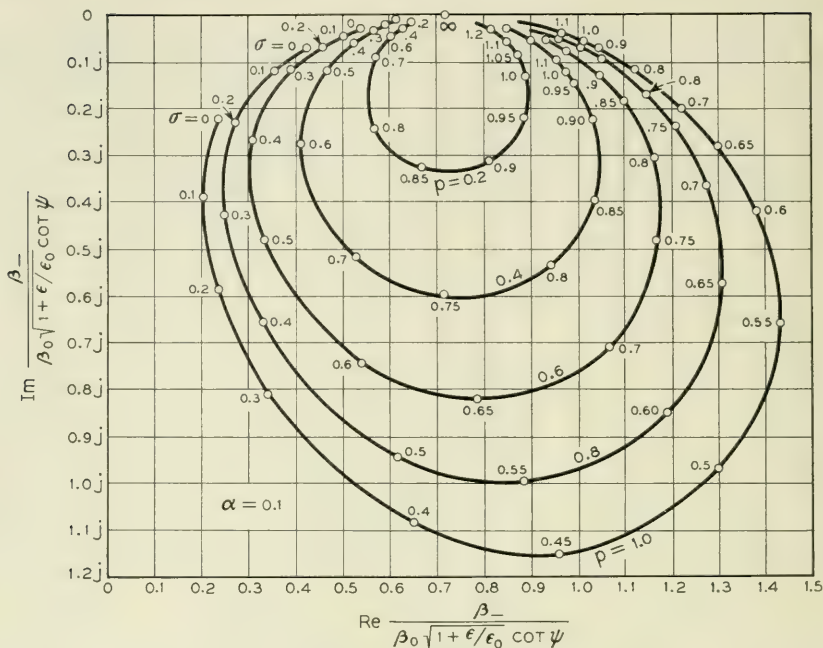


Fig. 6 — Real and imaginary parts of the reverse propagation constant β_- of a flat helix versus σ for various p and for $\alpha = 0.1$ (the parameter σ is marked along the curves). The forward propagation constant has a very small imaginary part which hardly varies with σ .

are shown in Figs. 9(a), (b) and (c), where β is plotted against x_0 for various fixed σ for two values of p . It is to be noted that the introduction of any characteristic length or scale into the problem, such as is provided here by the distance x_0 immediately produces a great complication in the mode spectrum. The plane helix with the ferrite in contact may be thought of as a highly degenerate problem.

To carry out loss calculations using the appropriate expressions for $\mu + \kappa \operatorname{sgn} \beta$ would be very tedious in the separated case. However, it was pointed out in Part 1 that to order α the expressions for μ and κ are given correctly if we put $\sigma + j\alpha$ in place of σ in the lossless formulae and that, in consequence, for small α , the imaginary part of the propagation constant is approximately given by

$$\alpha \frac{\partial \beta}{\partial \sigma}$$

In Figs. 10(a) and 10(b) the loss calculated in this way for the cases considered in Figs. 9 is shown.

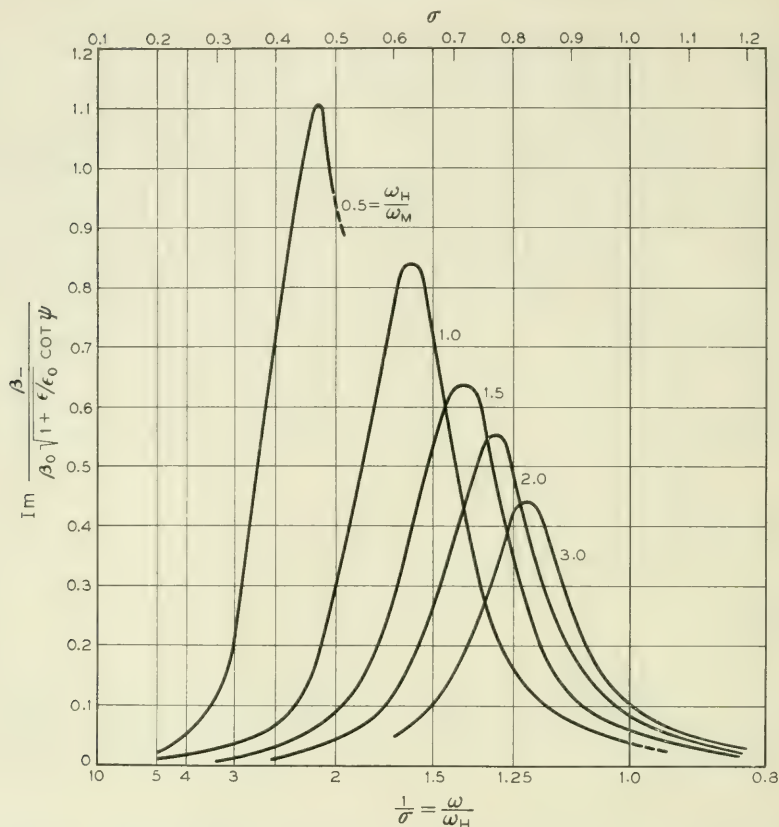


Fig. 8 — Attenuation of reverse wave versus frequency at a fixed magnetic field and saturation magnetization for various values of $\omega_H/\omega_M = (\mu_0 \bar{H})/M$. The curves for 0.5 and 1 are discontinued when p reaches unity.

are supposed to be independent of radial distance from the cylinder axis. For the geometries employed in practice, this will be a reasonably good assumption. Thus it is possible to relate the components of B and H in cylindrical coordinates (r, φ, z) through the tensor

$$\begin{pmatrix} \mu & 0 & -j\kappa \\ 0 & \mu_0 & 0 \\ j\kappa & 0 & \mu \end{pmatrix},$$

where μ and κ are given by the Polder relations (1) and are independent of position.

Written in cylindrical coordinates, Maxwell's equations in the ferrite are therefore

$$j\beta H_\varphi = j\omega\epsilon_1 E_r, \quad (20a)$$

$$-j\beta H_r - \frac{\partial H_z}{\partial r} = j\omega\epsilon_1 E_\varphi, \quad (20b)$$

$$\frac{1}{r} \frac{\partial}{\partial r} r H_\varphi = j\omega\epsilon_1 E_z, \quad (20c)$$

$$j\beta E_\varphi = -j\omega\mu(H_r - j\rho_H H_z), \quad (20d)$$

$$-j\beta E_r - \frac{\partial E_z}{\partial r} = -j\omega\mu_0 H_\varphi, \quad (20e)$$

$$\frac{1}{r} \frac{\partial}{\partial r} r E_\varphi = -j\omega\mu(j\rho_H H_r + H_z). \quad (20f)$$

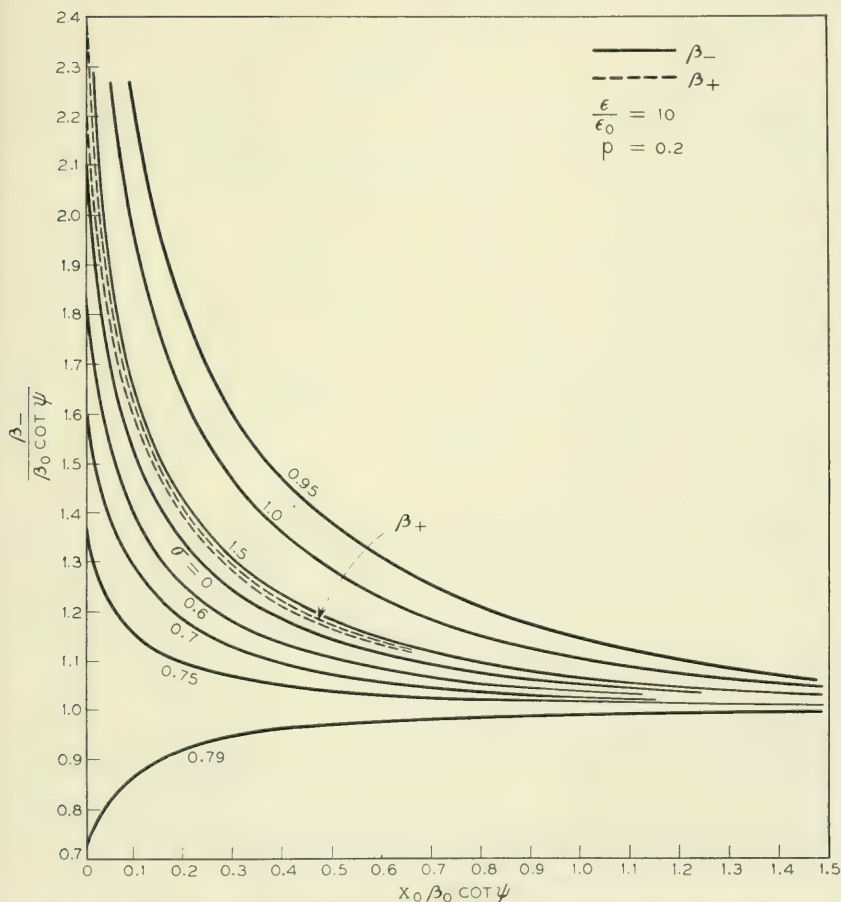


Fig. 9 — Propagation constants for a flat helix separated from the ferrite by a distance x_0 for various values of σ . (Loss free case) (a) β_- and β_+ for $p = 0.2$, (b) β_- for $p = 0.8$, (c) β_+ for $p = 0.8$. In (a), above, the dotted lines bound all σ values.

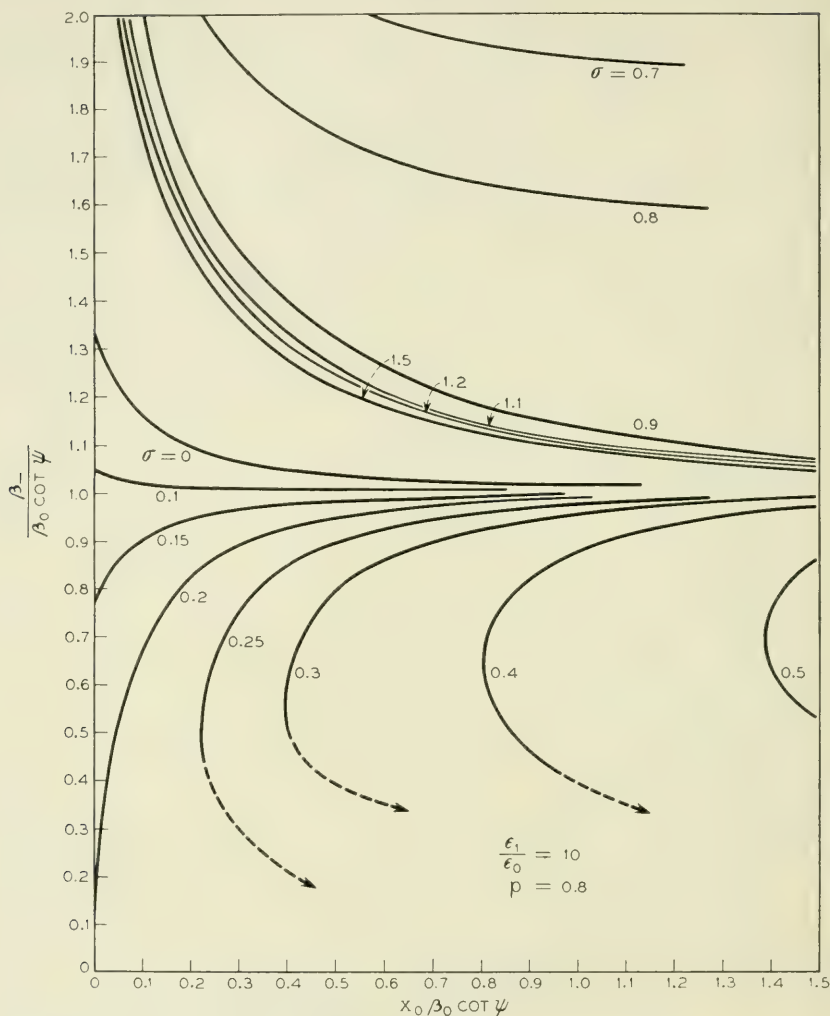


Fig. 9(b) — See Fig. 9.

As in the case of planar geometry, the field-components can be grouped into TE and TM parts; only the TE part will depend explicitly on the gyromagnetic properties of the medium. Equations (20a, 20c) and (20e) determine the TM field. E_r and H_φ can be eliminated from them, yielding the familiar wave equation

$$\frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial E_z}{\partial r} + (\beta_1^2 - \beta^2) E_z = 0; \quad \beta_1^2 = \omega^2 \mu_0 \epsilon_1, \quad (21)$$

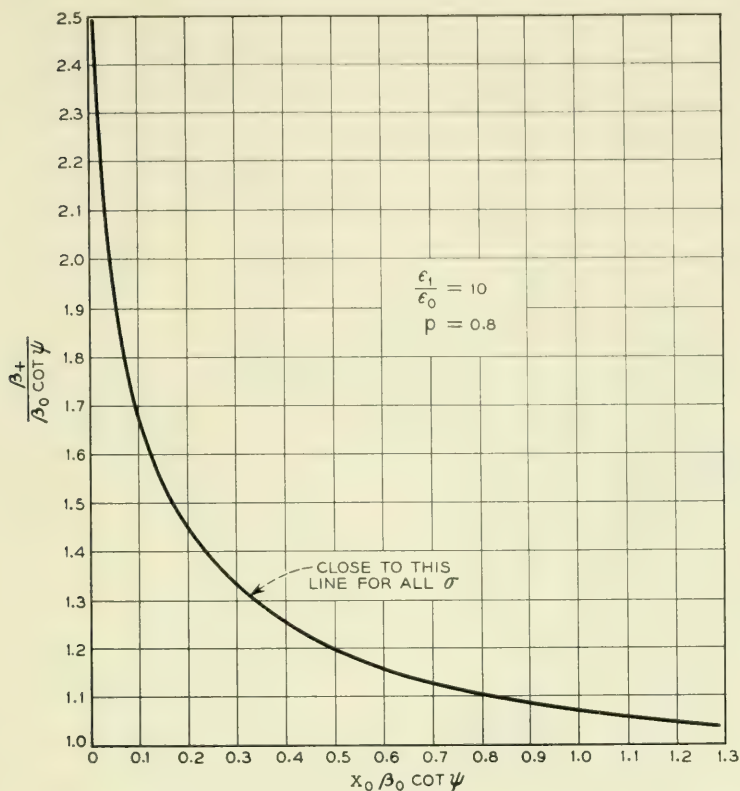


Fig. 9(c) — See Fig. 9.

whose solutions are zero order Bessel functions (or linear combinations thereof) of a kind depending on the region under consideration. Thus if $\beta > \beta_1$, and the region occupied by the medium includes the cylinder axis, the modified Bessel function $I_0(r\sqrt{\beta^2 - \beta_1^2})$ has to be chosen if the field is to be finite at $r = 0$; if the medium extends from a finite r to infinity, the function $K_0(r\sqrt{\beta^2 - \beta_1^2})$, regular at infinity, is selected. Correspondingly, if $\beta < \beta_1$, the Bessel and Hankel function $J_0(r\sqrt{\beta_1^2 - \beta^2})$ and $H_0^{(1,2)}(r\sqrt{\beta_1^2 - \beta^2})$ replace I_0 and K_0 respectively.

In terms of the appropriate solution of (21), the remaining field components are

$$E_r = -j \frac{\beta}{\beta_1^2 - \beta^2} \frac{\partial E_z}{\partial r}; \quad H_\varphi = -\frac{j\omega\epsilon_1}{\beta_1^2 - \beta^2} \frac{\partial E_z}{\partial r}.$$

The tangential admittance for the TM field is thus

$$Y_{\text{TM}} = \frac{H_\varphi}{E_z} = -\frac{j\omega\epsilon_1}{\beta_1^2 - \beta^2} \frac{\partial}{\partial r} \log E_z. \quad (22)$$

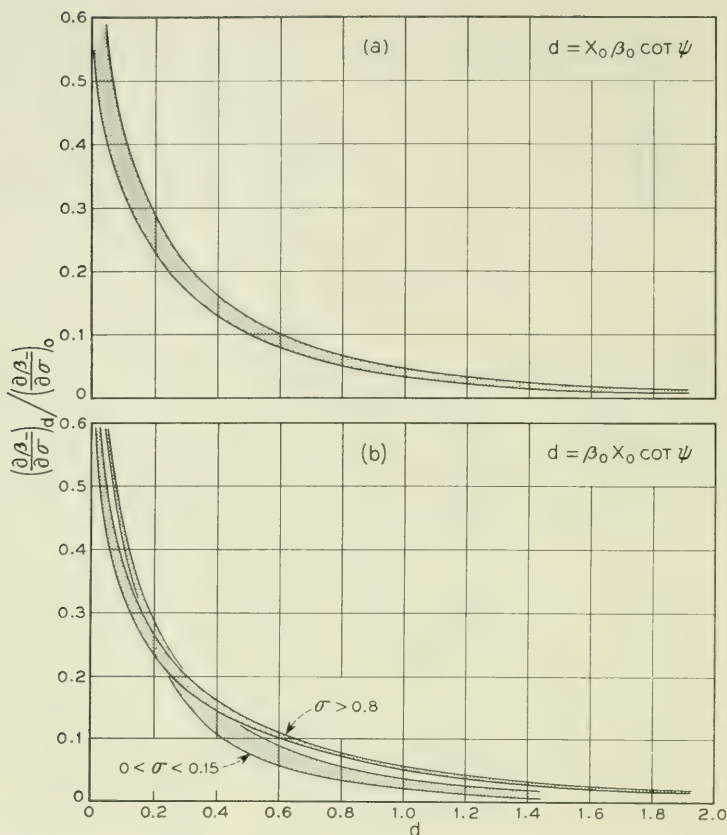


Fig. 10 — Ratios of reverse attenuation for a flat helix separated from the ferrite by a distance x_0 to the reverse attenuation at $x_0 = 0$. Computations made from the approximate slope-formula for the loss, where applicable. (a) $p = 0.2$. (b) $p = 0.8$. In (a) all applicable σ values lie in the shaded region.

For example, if $\beta > \beta_1$, and the medium occupies all space from a finite r to infinity

$$Y_{\text{TM}} = \frac{j\omega\epsilon_1 K'_0(\alpha_1 r)}{\alpha_1 K_0(\alpha_1 r)} = \frac{-j\omega\epsilon_1 K_1(\alpha_1 r)}{\alpha_1 K_0(\alpha_1 r)}, \quad (23)$$

where $\alpha_1 = \sqrt{\beta^2 - \beta_1^2}$.

The field components of the TE field are determined by equations (20b), (20d) and (20f). Elimination of E_φ and H_r from these gives

$$\frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial H_z}{\partial r} + \left(\beta_f^2 - \beta^2 - \frac{\rho_H \beta}{r} \right) H_z = 0, \quad (24)$$

where $\beta_f^2 = \omega^2 \mu \epsilon_1 (1 - \rho_H^2)$. The term $\beta_f^2 - \beta^2$ in the bracket is already

familiar from the planar case; it depends on the magnetic field and on the propagation direction in a purely reciprocal way. The term $\rho_H\beta/r$, however, reverses sign when either magnetizing field or propagation constant changes sign. The solutions of equation (24) are therefore different for opposite propagation directions (or opposite magnetizations). Thus, in contrast to the planar case, where non-reciprocity arose only through the boundary conditions, the cylindrical case is inherently non-reciprocal.

In the absence of the last term in the bracket, equation (24) would be solved by zero order Bessel functions, just like equation (21). In the presence of this term, the solutions become confluent hypergeometric functions. Different changes of variable bring these functions into forms known by different names and notations. One such change leads to Laguerre functions, another to Whittaker functions. We shall choose the latter representation, since it is closely related to Bessel functions, and numerical tables seem about equally scarce for all representations. In equation (24), let $\beta^2 - \beta_f^2 = \alpha_2^2$, and let $\alpha_2 r = y$. Then it becomes

$$\frac{1}{y} \frac{d}{dy} y \frac{dH_z}{dy} - \left(1 - \frac{2\chi}{y}\right) H_z = 0, \quad (25)$$

where $\chi = -\beta\rho_H/2\alpha_2$. Further, let $y = x/2$, and $H_z(y) = h(x)/\sqrt{x}$; then equation (25) becomes

$$\frac{d^2 h}{dx^2} + \left(\frac{1}{4x^2} + \frac{\chi}{x} - \frac{1}{4}\right) h = 0, \quad (26)$$

which is the standard form of the equation for zero order Whittaker functions. It is a special case of the equation for μ^{th} order Whittaker functions:

$$\frac{d^2 h}{dx^2} + \left(\frac{\frac{1}{4} - \mu^2}{x^2} + \frac{\chi}{x} - \frac{1}{4}\right) h = 0, \quad (27)$$

The solution of equation (27) which is regular near zero is denoted in the literature by $M_{\chi,\mu}(x)$; it is proportional, in the limit $\chi = 0$, to the Bessel function $I_\mu(x/2)$. The solution of equation (27) regular at infinity is denoted by $W_{\chi,\mu}(x)$, and in the limit $\chi = 0$, is proportional to $K_\mu(x/2)$. The factors of proportionality are found in Appendix I.

In this notation, the solutions for H_z are thus

$$\frac{M_{\chi,0}(2\alpha_2 r)}{\sqrt{2\alpha_2 r}}, \quad \frac{W_{\chi,0}(2\alpha_2 r)}{\sqrt{2\alpha_2 r}}^*.$$

* If $\beta < \beta_2$, both argument and suffix χ are imaginary. These functions are then related to J_0 and H_0 respectively.

Once the appropriate H_z is determined, H_r and E_φ are given by

$$\begin{aligned} H_r &= \frac{j}{\beta^2 - \omega^2 \mu \epsilon_1} \left(\beta \frac{\partial H_z}{\partial r} - \omega^2 \mu \epsilon_1 \rho_H H_z \right), \\ E_\varphi &= -\frac{j \omega \mu}{\beta^2 - \omega^2 \mu \epsilon_1} \left(\frac{\partial H_z}{\partial r} - \rho_H \beta H_z \right). \end{aligned} \quad (28)$$

The tangential impedance for TE-fields is thus

$$Z_{\text{TE}} = \frac{E_\varphi}{H_z} = \frac{\omega \mu}{j(\beta^2 - \omega^2 \mu \epsilon_1)} \left(\frac{\partial}{\partial r} \log H_z - \rho_H \beta \right). \quad (29)$$

The reader will recall that for isotropic media ($\rho_H = 0$) the numerator of the right hand side of equation (29) can always be expressed as the ratio of first order to zero order Bessel functions by virtue of relations like $K_0'(\chi) = -K_1(\chi)$, $I_0'(\chi) = I_1(\chi)$ and so on. Analogous results hold true in the present case. Suppose, for example, that we are dealing with a geometry such that the correct H_z is

$$H_z = \frac{W_{\chi,0}(2\alpha_2 r)}{\sqrt{2\alpha_2 r}} = R_{\chi,0}(2\alpha_2 r), \quad \text{say}$$

Then

$$\frac{1}{H_z} \frac{\partial H_z}{\partial r} = 2\alpha_2 R'_{\chi,0}/R_{\chi,0},$$

and

$$Z_{\text{TE}} = \frac{2\alpha_2 \omega \mu}{j(\beta^2 - \omega^2 \mu \epsilon_1)} \frac{R'_{\chi,0} + \chi R_{\chi,0}}{R_{\chi,0}}.$$

It is shown in Appendix I that

$$R_{\chi,0}' + \chi R_{\chi,0} = (\chi - \frac{1}{2}) R_{\chi,1}.$$

Therefore

$$\begin{aligned} Z_{\text{TE}} &= \frac{\omega \mu \alpha_2 (2\chi - 1)}{j(\beta^2 - \omega^2 \mu \epsilon_1)} \frac{R_{\chi,1}(2\alpha_2 r)}{R_{\chi,0}(2\alpha_2 r)}, \\ &= \frac{\omega \mu \alpha_2 (2\chi - 1)}{j(\beta^2 - \omega^2 \mu \epsilon_1)} \frac{W_{\chi,1}(2\alpha_2 r)}{W_{\chi,0}(2\alpha_2 r)}. \end{aligned} \quad (30)$$

A similar difference relation shows that if the region is such that

$$H_z = M_{\chi,0}(2\alpha_2 r)/\sqrt{2\alpha_2 r},$$

then

$$Z_{\text{TE}} = \frac{\omega\mu\alpha_2(1_4 - \chi^2)M_{\chi,1}(2\alpha_2r)}{j(\beta^2 - \omega^2\mu\epsilon_1)M_{\chi,0}(2\alpha_2r)}. \quad (31)^*$$

In the unmagnetized case equations (30) and (31) reduce to

$$-\frac{\omega\mu_0\alpha_1}{j(\beta^2 - \omega^2\mu_0\epsilon_1)} \frac{K_1(\alpha_1r)}{K_0(\alpha_1r)} = -\frac{\omega\mu_0K_1(\alpha_1r)}{j\alpha_1K_0(\alpha_1r)}, \quad (32)$$

and to

$$\frac{\omega\mu_0}{j\alpha_1} \frac{I_1(\alpha_1r)}{I_0(\alpha_1r)}. \quad (33)$$

One might be led to believe that the search for solutions of Maxwell's equations with angular dependence $e^{jn\varphi}$ will lead to n^{th} order Whittaker functions (just as in the isotropic case this dependence leads to n^{th} order Bessel functions). Such is not the case, however. Unless $n = 0$, one is led to two simultaneous second order equations for E_z and H_z , and the character of the problem is changed completely.

3.2. The Cylindrical Helix

We are now in a position to derive the characteristic equation for a close wound cylindrical helix and approximated by a helically conducting sheet surrounded by ferrite. We confine the discussion to the case in which the ferrite is in actual contact with the helix, Fig. 11; the case of finite separation discussed for the planar helix (Section 2.3) would be too cumbersome here. Losses are neglected. If they are not excessive, they can be deduced from the curves for the propagation constant in the loss free sample by differentiation, as outlined in Sections 2.1 and 4.15, Part I.

The boundary conditions are just the same as in the planar case. In Section 2.3 they were stated in terms of admittances, and it is only necessary to substitute for these the admittances just derived for cylindrical geometry. Thus for H_y/E_z we substitute Y_{TM} , and for H_z/E_φ we write $Y_{\text{TE}} = 1/Z_{\text{TE}}$.

If superfixes i and e refer to the inside and outside of the helix (in Section 2.3 on the plane helix i and e were denoted by “-”, and “+”), the characteristic equation is

$$\{Y_{\text{TM}}^{(i)} - Y_{\text{TM}}^{(e)}\}_{r=r_0} \cot^2\psi = (Y_{\text{TE}}^{(i)} - Y_{\text{TE}}^{(e)})_{r=r_0}, \quad (34)$$

where r_0 is the radius of the helix.

* The appearance of different factors $(2\chi - 1)$ and $(1_4 - \chi^2)$ is simply due to the way the functions W, M are normalized in the literature.

For waves bound to the helix, $Y_{TE}^{(e)}$ is to be derived from that solution of (24) which tends to zero as $r \rightarrow \infty$. This solution is $W_{x,0}(2\alpha_2 r)/\sqrt{2\alpha_2 r}$, and so $Y_{TE}^{(e)}$ is given by equation (30)

$$Y_{TE}^{(e)} = \frac{1}{Z_{TE}} = \frac{j}{\omega\mu} \frac{\beta^2 - \omega^2\mu\epsilon_1}{\alpha_2(2\chi - 1)} \frac{W_{x,0}(2\alpha_2 r)}{W_{x,1}(2\alpha_2 r)}.$$

Similarly, from equation (23), we have, for bound waves, with $E_z \sim K_0$,

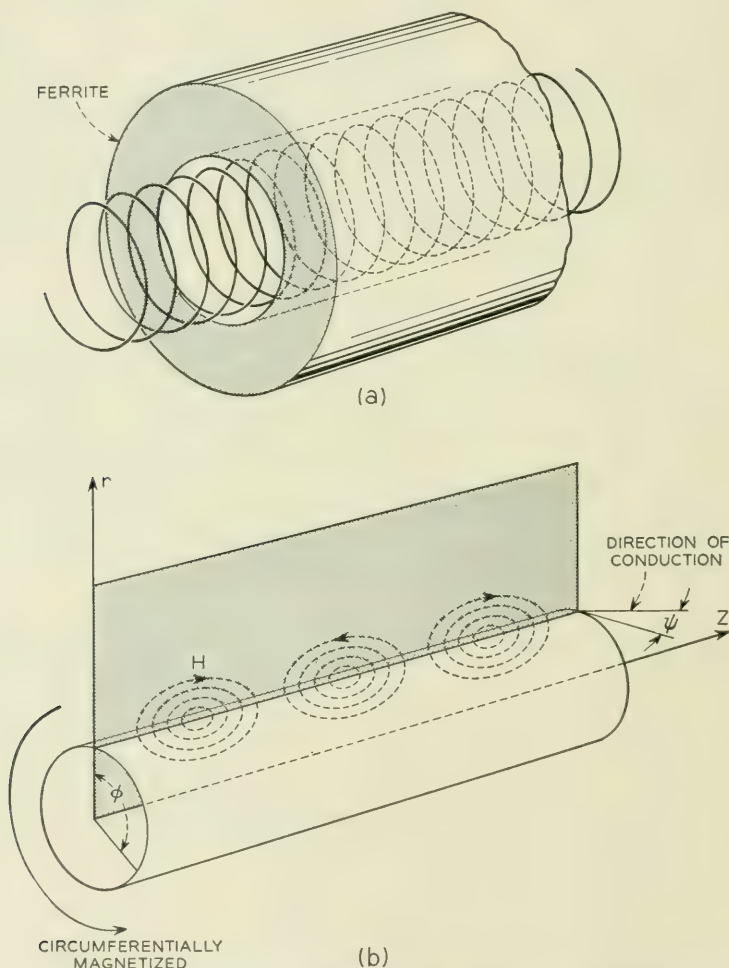


Fig. 11 — (a) Cylindrical helix surrounded by ferrite, (b) Magnetic field lines projected onto a plane containing the axis of the helix.

$$Y_{\text{TM}}^{(c)} = -\frac{j\omega\epsilon_1}{\alpha_1} \frac{K_1(\alpha_1 r)}{K_0(\alpha_1 r)}.$$

Inside the helix, we require the solution for an isotropic region which remains regular as $r \rightarrow 0$. Accordingly

$$Y_{\text{TE}}^{(i)} = \frac{j\alpha_0}{\omega\mu_0} \frac{I_0(\alpha_0 r)}{I_1(\alpha_0 r)}; \quad \alpha_0 = \sqrt{\beta^2 - \beta_0^2},$$

$$\beta_0^2 = \omega^2 \mu_0 \epsilon_0$$

and

$$Y_{\text{TM}}^{(i)} = j \frac{\omega\epsilon_0}{\alpha_0} \frac{I_1(\alpha_0 r_0)}{I_0(\alpha_0 r_0)},$$

where ϵ_0 , μ_0 are the dielectric constant, and permeability of vacuum. Combining these expressions in (34), we obtain after slight rearrangement

$$\frac{I_1(\alpha_0 r_0)}{I_0(\alpha_0 r_0)} + \frac{\alpha_0 \epsilon_1}{\alpha_1 \epsilon_0} \frac{K_1(\alpha_1 r_0)}{K_0(\alpha_1 r_0)} = \left[\alpha_0^2 \frac{I_0(\alpha_0 r_0)}{I_1(\alpha_0 r_0)} - \frac{\mu_0 \alpha_0 \beta^2 - \omega^2 \mu \epsilon_1}{\mu \alpha_2 (2\chi - 1)} \frac{W_{\chi,0}(2\alpha_2 r)}{W_{\chi,1}(2\alpha_2 r)} \right] \tan^2 \psi \quad (35)$$

which determines β .

A complete solution of equation (35) is out of the question. However, as in the planar case, for the slow waves used in traveling wave tube work, the equation may be simplified so that solutions may be computed rather easily. For electron velocities usually employed the resultant β must be about $10\beta_0$. Therefore in equation (35) it will be permissible to neglect all the quantities β_0^2 , β_1^2 , β_2^2 , $\omega^2 \mu \epsilon_1$, in comparison with β^2 , except in the narrow ranges of magnetic field such that μ or $\mu(1 - \rho_H^2)$ becomes very large. This will occur near $\pm\sigma_0$ where $\sigma_0 = -p/2 + \sqrt{p^2/4 + 1}$, and near $\sigma = 1$. A solution obtained by assuming a large β must be self-consistent; that is, it can be credited only in regions where it does, in fact, predict large β . However, in Section 2.3 it was shown for the plane helix that in any practical case the ranges of magnetic fields so excluded are very narrow, even in the loss-free case, and one may suppose that this is true also in the cylindrical case.

For slow waves, each of the α 's reduces to $|\beta|$; the absolute value sign derives from the fact that the positive square root is implied in the definition of the α 's. Therefore

$$\chi \rightarrow \frac{-\rho_H \beta}{2|\beta|} = -\frac{\rho_H}{2} \operatorname{sgn} \beta. \quad (36)$$

Now the suffix χ of the Whittaker functions no longer depends on the

magnitude of β , and it is chiefly for this reason that further progress is possible. For large β , equation (35) can be written

$$\beta^2 = \beta_0^2 \cot^2 \psi \frac{\frac{I_1(|\beta| r_0)}{I_0(|\beta| r_0)} + \frac{\epsilon_1}{\epsilon_0} \frac{K_1(|\beta| r_0)}{K_0(|\beta| r_0)}}{\frac{I_1(|\beta| r_0)}{I_0(|\beta| r_0)} - \frac{1}{\frac{\mu}{\mu_0} (2\chi - 1)} \frac{W_{\chi,0}(2|\beta| r_0)}{W_{\chi,0}(2|\beta| r_0)}} \quad (37)$$

where χ is now given by equation (36). Equation (37) is now solved by

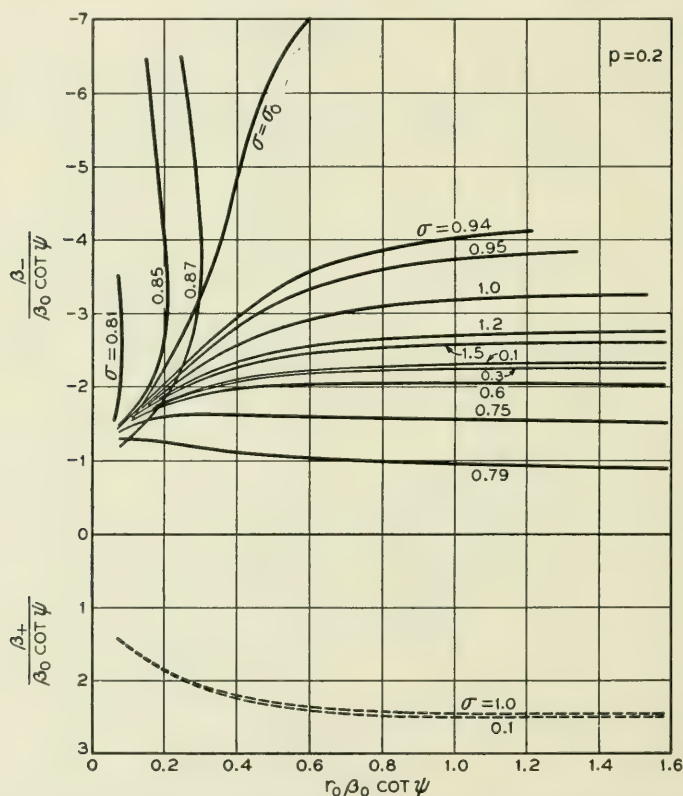


Fig. 12 — Reduced forward and reverse propagation constants versus reduced radius of a cylindrical helix (loss-free case) for various σ and p . The range $\sigma_1 < \sigma < \sigma_0$ where

$$\sigma_1 = \sqrt{1 + \frac{p^2}{4} - \frac{p}{3} - \frac{p}{2}} \quad \text{and} \quad \sigma_0 = \sqrt{1 + \frac{p^2}{4} - \frac{p}{2}}$$

contains an infinity of shape resonances and is not shown here. (a), above, $p = 0.2$. (b) $p = 0.6$. (c) $p = 1.0$.

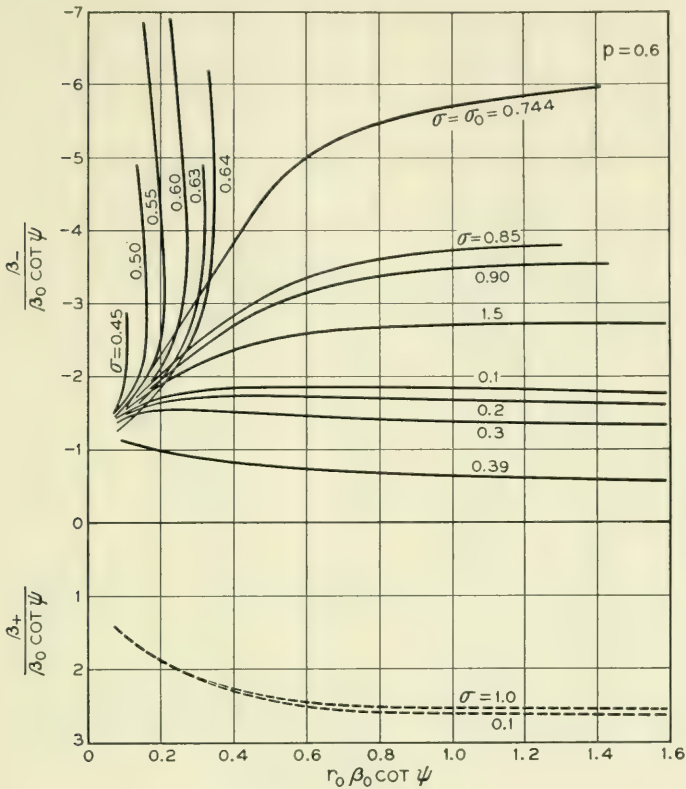


Fig. 12(b) — See Fig. 12.

the following procedure: Introduce the parameter

$$\frac{u}{2} = |\beta| r_0. \tag{38}$$

For a given ρ_H and u (or σ and p), and a given sign of β , each value of u determines β through equation (37), and then r_0 through equation (38). Thus β can be plotted versus r_0 . The procedure is repeated for the opposite sign of β (and therefore the opposite sign of χ). A different curve of β versus r_0 is then obtained. Thus for a given value of r_0 , the “forward” and “backward” propagation constants are different in magnitude. The results (computed for a typical ratio $\epsilon_1/\epsilon_0 = 10$) are conveniently stated in terms of $\bar{\beta} = \beta/(\beta_0 \cot \psi)$ and $\bar{r}_0 = r_0 \beta_0 \cot \psi$ and are shown in Fig. 12(a) to (c), and again, for fixed $\bar{r}_0$, in Fig. 13(a) to (e). We note that for $\bar{r}_0$ in excess of about 1.5, the results are almost the same as those

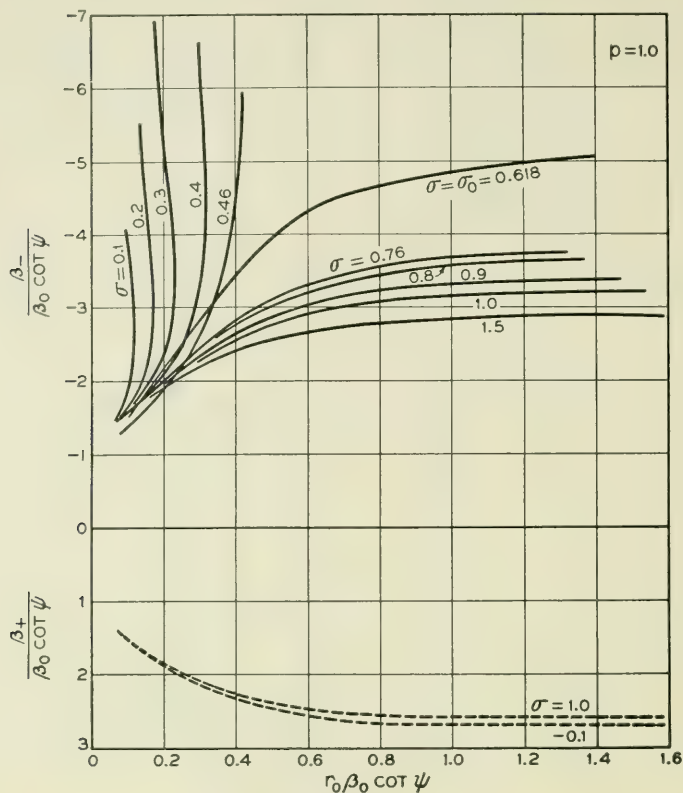


Fig. 12(c) — See Fig. 12.

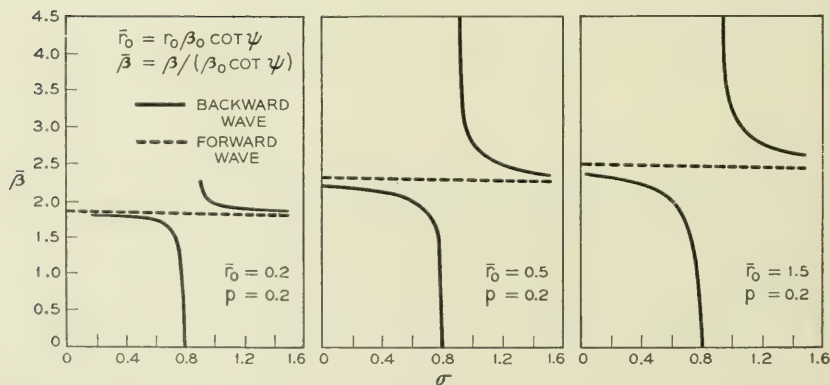


Fig. 13 — Reduced forward and reverse propagation constants versus σ for various reduced radii of a cylindrical helix. (Loss free case). The region $1 - p < \sigma < \sigma_0$ is omitted. (a), above, $p = 0.2$. (b) $p = 0.4$. (c) $p = 0.6$. (d) $p = 0.8$. (e) $p = 1.0$.

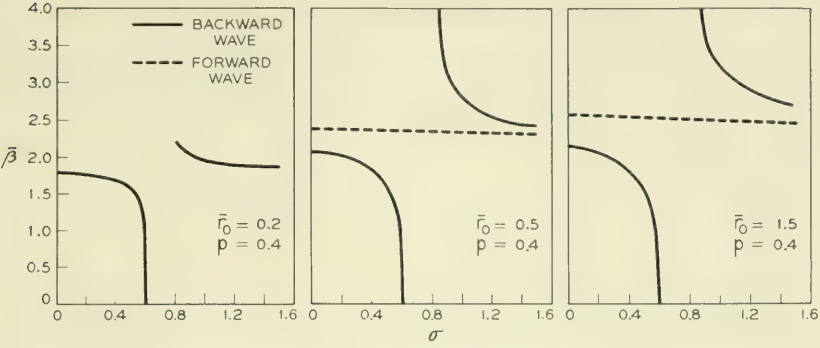


Fig. 13(b) — See Fig. 13.

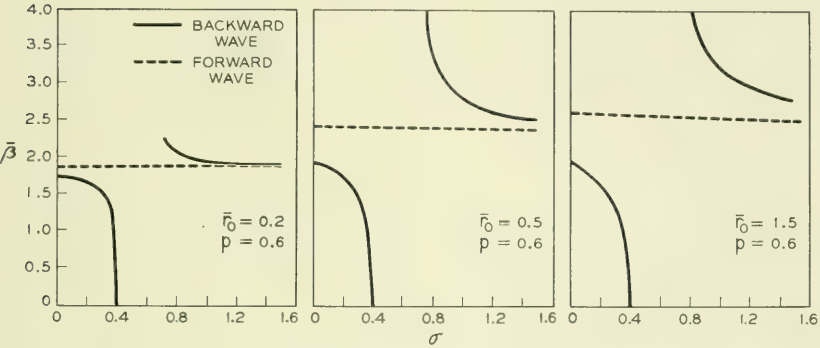


Fig. 13(c) — See Fig. 13.

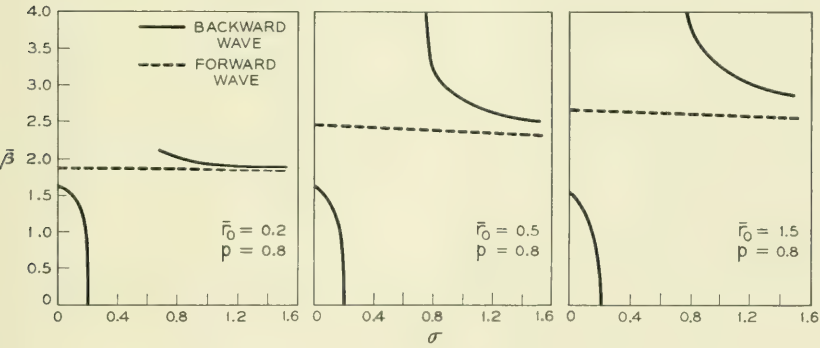


Fig. 13(d) — See Fig. 13.

for a flat helix. This is due to the fact that the large dielectric constant reduces the circuit wavelength so much that the radius appears infinite by comparison once it exceeds 1.5. In traveling-wave tube practice, however, $\bar{r}_0$ is generally below 1.5.

The behavior of the $\bar{\beta} - \bar{r}_0$ curves can be understood from the behavior of the ratio

$$\frac{W_{\chi,0}(u)}{W_{\chi,1}(u)} = Z_{\chi}(u)$$

and of the coefficient $1/[\mu/\mu_0 (2\chi - 1)]$. Suppose first that χ is positive. When χ exceeds zero only slightly, $Z_{\chi}(u)$ behaves essentially like $K_0(u/2)/K_1(u/2)$. This function is always positive; it begins at 0 when $u = 0$ with a vertical tangent and steadily increases to unity as $u \rightarrow \infty$. $Z_{\chi}(u)$ varies in the same way, see Figs. 14 and 15, in the range $0 < \chi < \frac{1}{2}$. For $\frac{3}{2} > \chi \geq \frac{1}{2}$, $W_{\chi,0}$, and therefore Z_{χ} , has a zero which increases from $u = 0$ at $\chi = \frac{1}{2}$ to $u = 1$ at $\chi = \frac{3}{2}$. Accordingly $Z_{\chi}(u)$ in $\frac{3}{2} >$

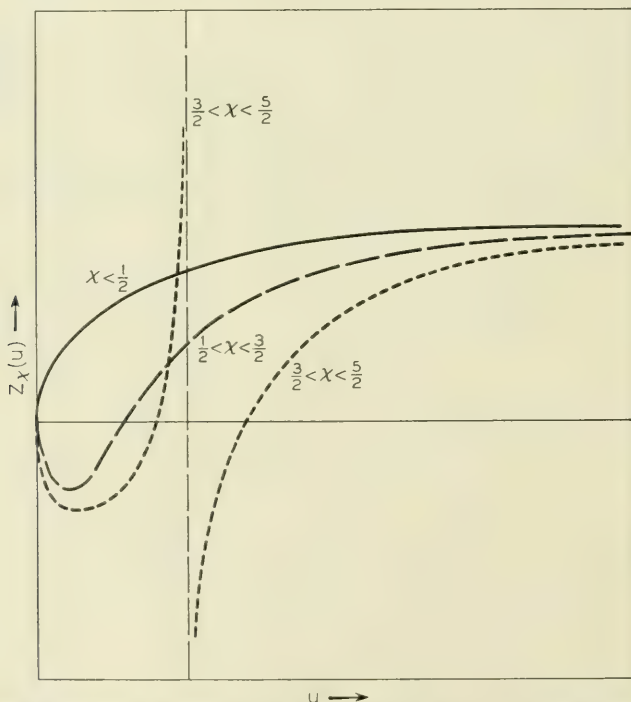


Fig. 14 — Schematic behavior of the function $Z_{\chi}(u) = W_{\chi,0}(u)/W_{\chi,1}(u)$ in various ranges of χ .

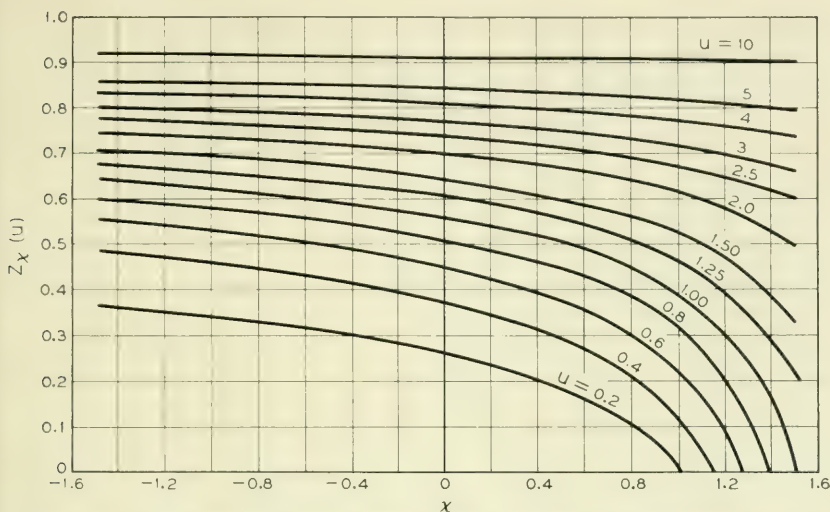


Fig. 15 — The function $Z_\chi(u)$ versus χ for various u , in the range $-\frac{3}{2} < \chi < \frac{3}{2}$.

$\chi > \frac{1}{2}$ starts from 0 at $u = 0$ with a downward-directed vertical tangent, achieves a negative minimum, then increases, through its zero, to the asymptotic value unity as $u \rightarrow \infty$. The minimum becomes deeper as χ approaches $\frac{3}{2}$. At $\chi = \frac{3}{2}$, $W_{\chi,1}(u)$ develops a zero at $u = 0$, which steadily moves to larger u as χ increases further towards $\frac{5}{2}$. At the same time the zero of $W_{\chi,0}$ already discussed moves from 1 to $2 + \sqrt{2}$, and a new zero arises at $u = 0$, $\chi = \frac{3}{2}$, which increases to $2 - \sqrt{2}$ as χ approaches $\frac{5}{2}$, but which lags behind the zero of $W_{\chi,1}(u)$. The function $Z_\chi(u)$ now has a pole and two zeros, and behaves as shown in Fig. 14. This process continues; each time χ passes $(2n + 1)/2$, a new zero and a new pole appear. (For a detailed list of poles and zeros the reader is referred to Appendix I). To apply these results, we first resort to the Polder relations.

In terms of σ , p , we have, for β negative

$$\chi = \frac{1}{2} \frac{p}{1 - p\sigma - \sigma^2}$$

and

$$\frac{1}{\frac{\mu}{\mu_0} (2\chi - 1)} = \frac{1 - \sigma}{p - 1 + \sigma} = A, \quad \text{say.}$$

The characteristic equation is

$$\bar{\beta}^2 = \frac{\frac{I_1(u/2)}{I_0(u/2)} + \frac{\epsilon_1}{\epsilon_0} \frac{K_1(u/2)}{K_0(u/2)}}{\frac{I_0(u/2)}{I_1(u/2)} - AZ_\chi(u)}; \quad |\bar{\beta}| \bar{r}_0 = u/2 \quad (37)$$

and can now be discussed in terms of σ at a fixed p . Suppose that $p < 1$. Then A is negative for $\sigma < 1 - p$. In the same range, $\chi < \frac{1}{2}$, so that Z behaves essentially as K_0/K_1 . Therefore both numerator and denominator are positive for all u , and the ratio tends to the planar result

$$\bar{\beta}^2 = \frac{1 + \frac{\epsilon_1}{\epsilon_0}}{1 - A}$$

as $u \rightarrow \infty$. As $u \rightarrow 0$, $\bar{\beta}^2$ tends to zero along the vertical, as can be shown by an examination of the various functions near $u = 0$. For $\sigma < 1 - p$, the course of the $\bar{\beta}^2$ versus u -curves is as shown schematically in Fig. 16(a), and it is easily seen that the $\bar{\beta}$ versus $\bar{r}_0$ curves run in essentially the same way, Fig. 12(a) to (c). However, as σ approaches $1 - p$, the $\bar{\beta}^2$ versus u curves steadily fall, until at $\sigma = 1 - p$, $\bar{\beta}^2 = 0$ for all finite u , since $A = -\infty$.

As σ passes $1 - p$, A changes sign and at the same time χ passes $\frac{1}{2}$ so that Z_χ acquires a zero. As σ varies from $1 - p$ to $1 - p/2$, A decreases from $+\infty$ to unity. Therefore, while $u < u_1$, the zero of Z_χ , $1 - AZ_\chi$ is positive; however as u increases beyond u_1 , $\{I_0(u/2)/I_1(u/2)\} - AZ_\chi(u)$ eventually passes zero, since $Z_\chi(u)$ and $I_0(u/2)/I_1(u/2)$ both approach unity. On the other hand the numerator of equation (37) is

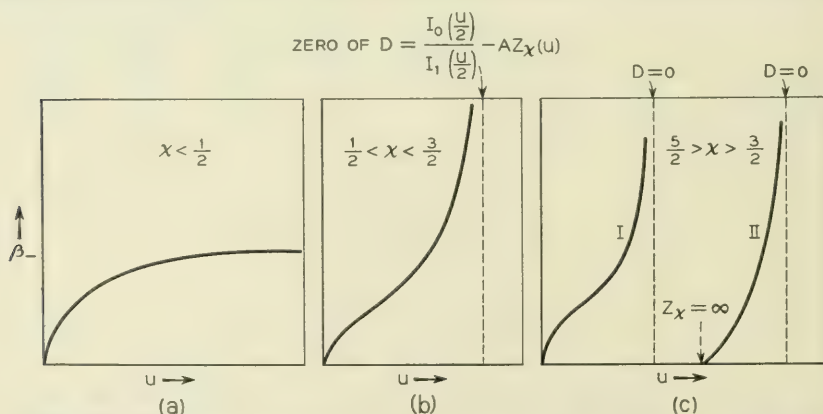


Fig. 16 — Schematic variation of β_- with u . a) $\chi < \frac{1}{2}$; b) $\frac{1}{2} < \chi < \frac{3}{2}$; c) $\frac{3}{2} < \chi < \frac{5}{2}$.

always positive; therefore $\bar{\beta}^2$ approaches infinity at $I_0(u/2)/I_1(u/2) - AZ_\chi(u) = 0$, and no real values of $\bar{\beta}$ exist thereafter (see Fig. 16b). Since this "cut-off" occurs at a finite value of u , the corresponding value of r_0 is zero. This explains the bulging of the corresponding $\bar{\beta} - \bar{r}_0$ curves in Fig. 12(a) to 12(c).

The next major change in the curves occurs when χ exceeds $\frac{3}{2}$, (that is, σ exceeds

$$\sigma_1 = -\frac{p}{2} + \sqrt{1 + \frac{p^2}{4} - \frac{p}{3}}.)$$

For $p < 2$, σ_1 is still less than $1 - (p/2)$, so that, initially at any rate, A is still greater than unity. In addition to the infinity of β^2 just discussed, a further infinity arises between $u = 0$, and the pole of $Z_\chi(u)$, as is seen from Fig. 16(c). β^2 increases from zero at $u = 0$ to this infinity, thereafter it is negative, until the pole of $Z_\chi(u)$ is reached. Thereupon it resumes at $\beta^2 = 0$ and approaches infinity at the zero of the denominator $[I_0(u/2)/I_1(u/2)] - AZ_\chi(u)$ already discussed. Thus there are now two branches of the $\bar{\beta}^2 - u$ curve; their corresponding traces in the $\bar{\beta} - \bar{r}_0$ plane are shown schematically in Fig. 17. (The computations on which Fig. 12(a) to 12(c) were based were broken off at $\sigma = \sigma_1$.)

A further branch is added each time χ increases beyond a number of the form $(2n + 1)/2$ (σ increases beyond

$$\sigma_n = -\frac{p}{2} + \sqrt{1 + \frac{p^2}{4} - \frac{p}{2n + 1}}.)$$

These all resemble the two branches just discussed, until $n >$

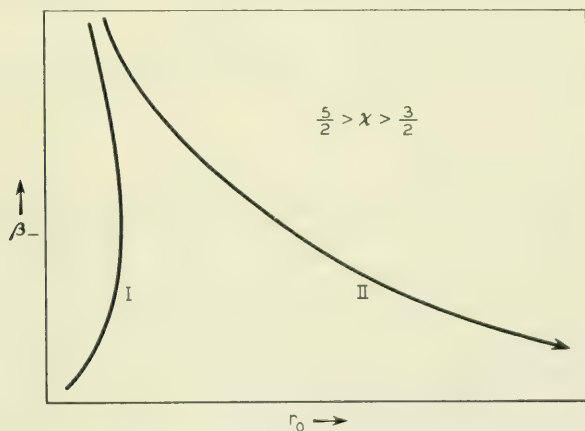


Fig. 17 — Schematic variation of β_- with r_0 for $\frac{3}{2} < \chi < \frac{5}{2}$.

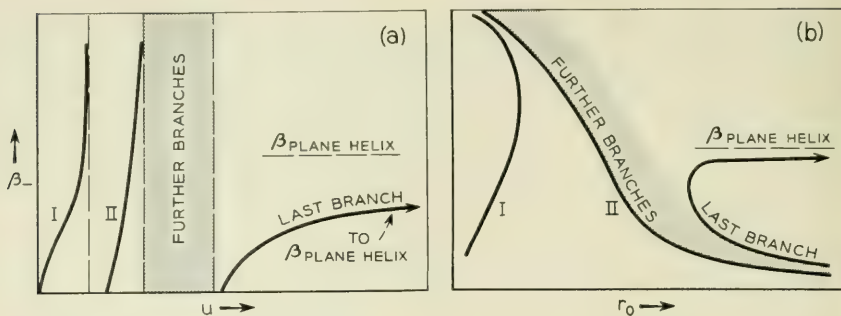


Fig. 18 — (a) β_- versus u when $1 - (p/2) < \sigma < \sigma_0$. (b) β_- versus r_0 under the same circumstances.

$\frac{1}{2} \left(\frac{4}{p} - 1 \right)$. When this occurs, $\sigma_n > 1 - (p/2)$, so that there will be a value σ between σ_{n-1} and σ_n beyond which the denominator no longer decreases through zero as $u \rightarrow \infty$, but approaches a finite positive value, Fig. 18(a). Accordingly β^2 approaches a finite positive value, and cut-off of the extreme right hand branch, Fig. 16(c), no longer occurs. The corresponding $\bar{\beta}$ versus $\bar{r}_0$ branch is as in Fig. 18(b).

As $n \rightarrow \infty$ ($\sigma_n \rightarrow \sigma_0 = \sqrt{1 + (p^2/4)} - p/2$) the number of branches increases to infinity. This situation resembles that in the completely filled waveguide (Part I), where we found an infinity of modes ("Shape-resonances") in the range $\sigma_0 < \sigma < 1$. In the present case, however, they are to be found in the range $1 - p < \sigma < \sigma_0$.

When $\sigma = \sigma_0 + 0$, χ is infinite and negative. The function $Z_\chi(u)$ is then constant and equal to unity. A is less than unity, and the denominator of equation (27) has no zeros. The $\bar{\beta}$ versus $\bar{r}_0$ curve is now "normal" again, see Fig. 12(a). As σ increases further, the curve falls (since A decreases steadily to -1 as $\sigma \rightarrow \infty$), and no more qualitative changes occur.

3.3 Cylindrical Waveguides

As pointed out before, the fact that the propagation problem in the cylindrical case can always be integrated in terms of Whittaker functions when the fields show no angular variation is an accident, and in view of the lack of numerical tables, not a particularly fortunate one. Only in special cases (like that of the slow-wave helix) is the text-book information on these functions of any great utility. In general, it will be more convenient to solve the differential equations numerically. However, for completeness, we shall state some of the formal results for a cylindrical waveguide containing a cylinder of circumferentially magnetized ferrite, and propagating a TE_0 mode.

First we consider a waveguide, radius r_0 , into which is fitted a hollow cylinder of ferrite, outer radius r_0 , inner radius r_1 . In that cylinder, the magnetic field H_z may be taken to be a superposition $AS_{\chi 0}(2\alpha_2 r) + BR_{\chi 0}(2\alpha_2 r)$ where, as before,

$$\alpha_2^2 = \beta^2 - \omega^2 \epsilon_1 \mu (1 - \rho_H^2); \quad S(x) = \frac{M(x)}{\sqrt{x}}; \quad R(x) = \frac{W(x)}{\sqrt{x}}.$$

$$\chi = -\frac{\beta \rho_H}{2\alpha_2}.$$

α_2 may be either real or positive imaginary. [In choosing this combination we depart from the usual practice of taking a superposition of J_0 and N_0 in the isotropic case. Were we to follow this practice, it would be necessary to define a new function $R_{\chi\mu}(2jx)e^{j(\mu\pi/2)} + R_{\chi\mu}(-2jx)e^{-j\mu\pi}$ to correspond to $N_\mu(x)$. Our choice corresponds to taking a combination of $J_0(x)$ and one of the Hankel functions $H_0(x)$ in the isotropic case. Since the functions H , J , N are linearly dependent, this will not affect the results.]

In view of the difference relations, equation (39) in Appendix I, and of equation (29) we obtain for the impedance in the ferrite

$$\frac{E_\varphi}{H_z} = -\frac{j\omega\mu\alpha_2}{(\beta^2 - \omega^2\mu\epsilon_1)} \frac{[A(1/4 - \chi^2)S_{\chi 1}(2\alpha_2 r) + B(2\chi - 1)R_{\chi 1}(2\alpha_2 r)]}{[AS_{\chi 0}(2\alpha_2 r) + BR_{\chi 0}(2\alpha_2 r)]}.$$

A and B must be adjusted so that this quantity vanishes at r_0 , the guide wall. This gives

$$\frac{E_\varphi}{H_z} = -(2\chi - 1)(1/4 - \chi^2) \frac{j\omega\mu\alpha_2}{\beta^2 - \omega^2\mu\epsilon_1} \frac{W_{\chi,1}(2\alpha_2 r_0)M_{\chi,1}(2\alpha_2 r) - M_{\chi,1}(2\alpha_2 r_0)W_{\chi,1}(2\alpha_2 r)}{(2\chi - 1)W_{\chi,1}(2\alpha_2 r_0)M_{\chi,0}(2\alpha_2 r) - (1/4 - \chi^2)M_{\chi,1}(2\alpha_2 r_0)W_{\chi,0}(2\alpha_2 r)}.$$

In the vacuum, between $r = 0$ and $r = r_1$, H_z is $I_0(\alpha_0 r)$ and the impedance is $E_\varphi/H_z = -j(\omega\mu_0/\alpha_0)I_1(\alpha_0 r)/I_0(\alpha_0 r)$ where $\alpha_0^2 = \beta^2 - \omega^2\mu\epsilon_0$. At $r = r_1$, the ferrite-vacuum interface, the two impedances must be equal. Thus we obtain the characteristic equation

$$\frac{\mu_0}{\alpha_0} \frac{I_1(\alpha_0 r_1)}{I_0(\alpha_0 r_1)} = (2\chi - 1)(1/4 - \chi^2) \frac{\alpha_2 \mu}{\beta^2 - \omega^2\mu\epsilon_1} \frac{W_{\chi,1}(2\alpha_2 r_0)M_{\chi,1}(2\alpha_2 r_1) - M_{\chi,1}(2\alpha_2 r_0)W_{\chi,1}(2\alpha_2 r_1)}{(2\chi - 1)W_{\chi,1}(2\alpha_2 r_0)M_{\chi,0}(2\alpha_2 r_1) - (1/4 - \chi^2)M_{\chi,1}(2\alpha_2 r_0)W_{\chi,0}(2\alpha_2 r_1)}.$$

(It is understood that for "normal" waveguide propagation α_0 will be imaginary, and the I will be replaced by J). As a simple illustration we

consider the case in which the ferrite cylinder fills the waveguide completely. E_φ is then proportional to $M_{\chi 1}(2\alpha_2 r)$, and so β is determined from the condition

$$M_{\chi 1}(2\alpha_2 r_0) = 0.$$

When "normal" propagation prevails (β less than the natural propagation constant of the medium $\omega\sqrt{\epsilon_1\mu(1-\rho_H^2)}$) both α_2 and χ are imaginary, equal to $j\alpha_2'$ and $j\chi'$, say. Under these circumstances little is known about the zeros of M . However, it is possible to say something about the solution for large radial mode numbers. It follows from Erdélyi et al.² **1**, p. 278, formula (2), that for large argument

$$M_{j\chi', 1}(2j\alpha_2' r_0) = \text{const} \cdot \sin[\alpha_2' r_0 + \chi' \log \alpha_2' r_0 + \chi' \log 2 + \Phi(\chi') - \pi/4],$$

where $\Phi(\chi') = \arg \Gamma(\frac{3}{2} + j\chi')$. The zeros of this expression are at

$$\alpha_2' r_0 + \chi' \log \alpha_2' r_0 = \frac{5n\pi}{4} - \chi' \log 2 - \Phi(\chi'). \quad n = \text{a large integer}$$

This equation may be solved graphically by setting $\alpha_2' r_0 = u$, assigning values to β , ρ_H , u (and hence to χ' , α_2'). From a solution u one then finds

$$r_0 = u/\alpha_2'.$$

M also has zeros for real α_2 , χ , if χ is large enough. Thus the waveguide will support waves with a β^2 greater than $\omega^2\mu\epsilon_1(1-\rho_H^2)$. It is shown in Reference 2 (**1**, p. 289) that when χ is between $\frac{3}{2}$ and $\frac{5}{2}$, M has one zero, when χ is between $\frac{5}{2}$ and $\frac{7}{2}$, M has two zeros and so on. Suppose that ρ_H is negative, $= -|\rho_H|$. Then

$$\chi = \frac{\beta |\rho_H|}{\sqrt{\beta^2 - \beta_2^2}}.$$

For real positive β , this equation has a solution for β if $|\rho_H| < \chi$:

$$\beta = \frac{\chi\beta_2}{\sqrt{\chi^2 - \rho_H^2}}.$$

If $\frac{3}{2} < \chi < \frac{5}{2}$, M will have a zero $u(\chi)$ depending on the value of χ . Thus the equations

$$\beta = \frac{\chi\beta_2}{\sqrt{\chi^2 - \rho_H^2}}; \quad r_0 = \frac{u(\chi)}{\sqrt{\beta^2 - \beta_2^2}} = \frac{\sqrt{\chi^2 - \rho_H^2}}{|\rho_H| \beta_2} u(\chi)$$

solve the propagation problem parametrically. Similarly when χ is between $\frac{5}{2}$ and $\frac{7}{2}$, there are two zeros of M given by two functions

$u_1(\chi)$ and $u_2(\chi)$. There are now two possible modes, with the same restrictions on ρ_H . An additional mode arises each time χ is allowed to pass a number of the form $(2n + 1)/2$. It is to be noted that these modes are not confined to the resonance range. For β positive, they can exist in the range $\infty > \sigma > \sigma_0$ and in the range $-\sigma_0 < \sigma < 0$.

APPENDIX I. SOME PROPERTIES OF WHITTAKER FUNCTIONS USED IN THIS PAPER

I. RELATION TO BESSEL FUNCTIONS

$$Lt_{\chi \rightarrow 0} \frac{M_{\chi, \mu}(2jx)}{\sqrt{2jx}} = 2^{2\mu} \Gamma(\mu + 1) j^\mu J_\mu(x),$$

$$Lt_{\chi \rightarrow 0} \frac{M_{\chi, \mu}(2x)}{\sqrt{2x}} = 2^{2\mu} \Gamma(\mu + 1) I_\mu(x),$$

$$Lt_{\chi \rightarrow 0} \frac{W_{\chi, \mu}(2x)}{\sqrt{2x}} = \frac{K_\mu(x)}{\sqrt{\pi}},$$

$$Lt_{\chi \rightarrow 0} \frac{W_{\chi, \mu}(-2jx)}{\sqrt{-2jx}} = \frac{\sqrt{\pi}}{2} j^{\mu+1} H_\mu^{(1)}(x), \quad \text{and}$$

$$Lt_{\chi \rightarrow 0} \left[\frac{W_{\chi, \mu}(2jx)}{\sqrt{2jx}} e^{j(\mu\pi/2)} + \frac{W_{\chi, \mu}(-2jx)}{\sqrt{-2jx}} e^{-j(\mu\pi/2)} \right] = -\sqrt{\pi} N_\mu(x)$$

II. DIFFERENCE RELATIONS

The following results can be obtained either by reference to Erdélyi,² 1 pp. 258, 254, by differentiation and subsequent integration by parts of integrals such as

$$W_{\chi, \mu}(x) = \frac{x^{\mu+1/2} e^{-x/2}}{\Gamma(\mu + 1/2 - \chi)} \int_0^\infty e^{-x\tau} \tau^{\mu-\chi-1/2} (H(1 + \tau))^{\mu+\chi-1/2} d\tau$$

or by observing that combinations of the form

$$\sqrt{x} \frac{d}{dx} \frac{W_{\chi, 0}(x)}{\sqrt{x}} + \chi W_{\chi, 0}(x)$$

satisfy Whittaker's equation with $\mu = 1$. In the last mentioned method, the required constant multiplying the first order Whittaker function can be obtained by reference to the limiting behavior for small x . If $R_{\chi\mu} = W_{\chi, \mu}(x)/\sqrt{x}$ and $S_{\chi\mu} = M_{\chi, \mu}(x)/\sqrt{x}$ the results are

$$\begin{aligned} R_{\chi 0'} + \chi R_{\chi 0} &= (\chi - 1/2) R_{\chi 1}, \\ S_{\chi 0'} + \chi S_{\chi 0} &= 1/2 (1/4 - \chi^2) S_{\chi 1}. \end{aligned} \quad (39)$$

III. ZEROS

When $\chi = (2n + 1)/2$, $n = 1, 2, \dots$, $W_{\chi, \mu}/x^{\mu-1/2}$ reduces to a polynomial times a function of χ . This may be inferred from the asymptotic expansion,

$$W_{\chi, \mu} = e^{-(x/2)} x^{\chi} \left(1 + \sum_{n=1}^{\infty} \frac{[\mu^2 - (\chi - 1/2)^2][\mu^2 - (\chi - 3/2)^2] \cdots [\mu^2 - (\chi - n + 1/2)^2]}{n! x^n} \right),$$

which terminates if $\mu = 1$ and $\chi = n + 1/2$ ($n = 1, 2, \dots$), or from the fact that for these values of the suffixes, W reduces to the generalized Laguerre polynomial

$$L_{\chi-(3/2)}^{(2)}(x).$$

Similarly when $\chi = (2n + 1)/2$, $n = 0, 1, 2, \dots$, $W_{\chi, 0}$ reduces to the Laguerre polynomial

$$L_{\chi-(1/2)}(x).$$

The zeros at the critical values of χ are given in the following table

χ	Zeros of $W_{\chi, 0}$	Zeros of $W_{\chi, 1}$
$1/2$	0	None
$3/2$	0; 1	0
$5/2$	0; 0.586; 3.414	0; 3
$7/2$	0; 0.416; 2.294; 6.290	0; 6; 2
$9/2$	0; 0.323; 1.746; 4.537; 9.395	0; 1.517; 4.312; 9.171
$11/2$	0; 0.26356; 1.413; 3.596; 7.086; 12.641	0; 1.227; 3.413; 6.903; 12.458

Between $n + 1/2$ and $n + 3/2$, $W_{\chi, 0}$ has $n + 1$ zeros ($n = 0, 1, 2, \dots$) and $W_{\chi, 1}$ has n zeros ($n = 1, 2, \dots$).

The zeros of $M_{\chi, 1}$ coincide with those of $W_{\chi, 1}$ when $\chi = n + 1/2$, and at those values of χ only. The functions $M_{\chi, 1}$ and $W_{\chi, 1}$ then are proportional to each other.*

IV. THE RICCATI-EQUATION FOR THE IMPEDANCE FUNCTION $W_{\chi, 0}/W_{\chi, 1}$

The computations concerning the cylindrical helix required a study of the function

$$Z(u) = \frac{W_{\chi, 0}(u)}{W_{\chi, 1}(u)}.$$

* At these critical values of χ , the solution to the problem of the hollow cylinder of ferrite in the waveguide breaks down, since M and W are then not independent. A further independent solution must then be constructed.

We will show that $Z_\chi(u)$ satisfies a non-linear first order differential equation of Riccati-type. From the difference relations, equation (36), we have

$$\begin{aligned} (\chi - 1/2) \frac{W_{\chi,1}(u)}{W_{\chi,0}(u)} &= \frac{W_{\chi,0}'(u)}{W_{\chi,0}(u)} - \frac{1}{2u} + \chi \\ &= \frac{1}{g}, \quad \text{say} \end{aligned}$$

and from Whittaker's equation

$$\frac{d}{du} \left(\frac{W_{\chi,0}'}{W_{\chi,0}} \right) + \left(\frac{W_{\chi,0}'}{W_{\chi,0}} \right)^2 + \left(\frac{1}{4u^2} + \frac{\chi}{u} - \frac{1}{4} \right) = 0.$$

Therefore

$$\frac{d}{du} \frac{1}{g} + \frac{1}{g^2} + \frac{1}{g} \left(\frac{1}{u} - 2\chi \right) + (\chi^2 - 1/4) = 0,$$

or

$$\frac{dg}{du} = 1 + \left(\frac{1}{u} - 2\chi \right) g + (\chi^2 - 1/4) g^2.$$

Finally, let

$$Z = (\chi - 1/2) g(u) = \frac{W_{\chi,0}(u)}{W_{\chi,1}(u)}.$$

Then

$$\frac{dZ}{du} = (\chi - 1/2) + \left(\frac{1}{u} - 2\chi \right) Z + (\chi + 1/2) Z^2.$$

Since this equation is satisfied by $M_{\chi,0}(u)/M_{\chi,1}(u)$ as well as by $W_{\chi,0}(u)/W_{\chi,1}(u)$, a selection has to be made from all the possible solutions of this equation. We require the one which for large u approaches unity. But for large u the equation is

$$\frac{dZ}{du} = (1 - Z) [\chi - 1/2 - (\chi + 1/2) Z],$$

whose integral is

$$Z = \frac{(\chi - 1/2)Ae^u - 1}{(\chi + 1/2)Ae^u - 1}$$

For large u , therefore, the solution is either unity, when $A = 0$, or else $(\chi - 1/2)/(\chi + 1/2)$, $A \neq 0$. The solution with $A \neq 0$ corresponds to the M functions; that with $A = 0$ to the W -functions.

The case $A = 0$ was integrated on an analogue-computer, and the results are shown in Fig. 14(b). The computation was restricted to the range $|\chi| < \frac{3}{2}$. Beyond these values, the helix-problem was discussed only qualitatively.

REFERENCES

1. M. L. Kales, H. N. Chait, and N. G. Sakiotis, Letter to the Editor, *J. Appl. Phys.*, **24**, No. 6.
2. Erdélyi, Magnus, Oberhettinger and Tricomi, *Higher Transcendental Functions*, **I**, McGraw-Hill, 1953.
3. E. H. Turner, *I. R. E. Proc.*, **41**, p. 937, July, 1953.
4. J. S. Cook, R. Kompfner, and H. Suhl, Non-Reciprocal Loss in Traveling Wave Ferrite Attenuators, Letter to Editor, *I. R. E. Proc.*, to be published.

Theoretical Fundamentals of Pulse Transmission — II

By E. D. SUNDE

(Manuscript received September 23, 1953)

PART II.

12. Impulse Characteristics and Pulse Train Envelopes	987
13. Transmission Limitations in Symmetrical Systems	991
14. Transmission Limitations in Asymmetrical Sideband Systems	996
15. Double vs. Vestigial Sideband Systems	1004
16. Limitation on Channel Capacity by Characteristic Distortion	1007
Acknowledgements	1010
References	1010

Part I of this paper dealt with various idealized transmission characteristics and with methods of evaluating pulse distortion resulting from various system imperfections. In Part II the resultant transmission impairments or limitations on pulse transmission rates are discussed for systems with low-pass, symmetrical band-pass and asymmetrical band-pass characteristics, and a comparison made of the transmission performance of double and vestigial sideband systems. The limitation on channel capacity imposed by random imperfections in the transmission-frequency characteristic, as compared to random noise, is also discussed.

12. IMPULSE CHARACTERISTICS AND PULSE TRAIN ENVELOPES

In pulse modulation systems pulses are transmitted in various combinations to form pulse trains, and at the receiving end the envelope of the pulse train is sampled at regular intervals to determine the amplitudes of the transmitted pulses. As a result of pulse overlaps there may be appreciable distortion of the pulse train envelope, which may cause errors in reception or noise, depending on the type of system. To evaluate transmission impairments, or limitations imposed on transmission capacity to avoid excessive transmission impairments from pulse distortion, it is necessary to establish basic relations between the impulse characteristic of the system and the envelope of the received pulse train.

In Fig. 42 are shown three transmitted pulses of different peak amplitudes, A_{-1} , A_0 and A_1 , transmitted at intervals τ with the first and third

pulse overlapping into the middle pulse. The instantaneous amplitude of the received train at a time t_0 referred to the peak amplitude of the middle pulse is

$$\begin{aligned} W(t_0) &= A_{-1}P(t_0 - \tau) + A_0P(t_0) + A_1P(t_0 + \tau), \\ &= \sum_{n=-1}^1 A_n P(t_0 + n\tau). \end{aligned} \quad (12.01)$$

When the sequence of pulses transmitted at uniform intervals τ extends between $n = -\infty$ and ∞ , the instantaneous amplitude of the pulse train at time t_0 is

$$W(t_0) = \sum_{n=-\infty}^{\infty} A_n P(t_0 + n\tau). \quad (12.02)$$

The above equation gives the instantaneous value $W(t_0)$ for any selected combination of transmitted pulses. The transmitted pulses may have any value within certain limits, as when they represent signal samples in a pulse amplitude modulation system, or may assume two or

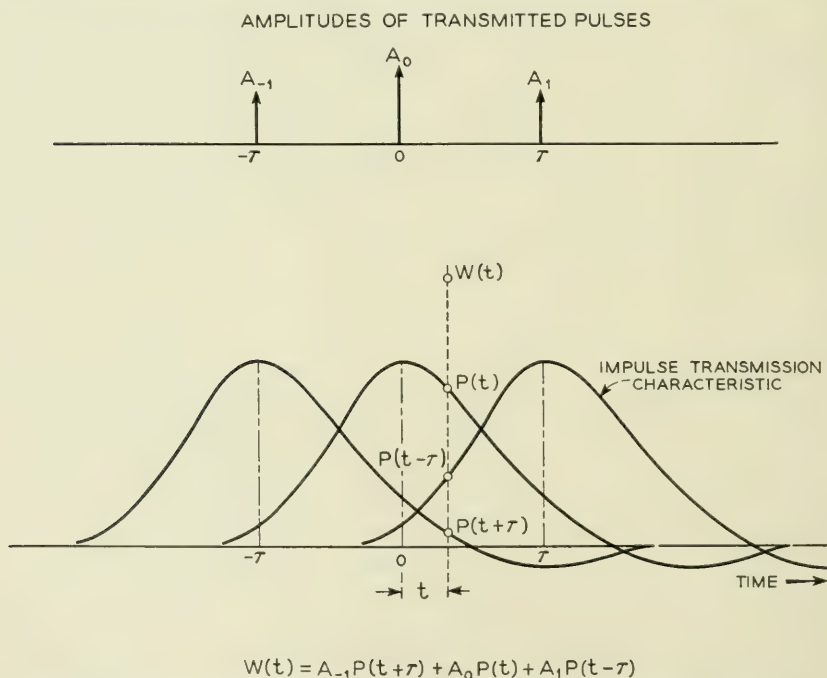


Fig. 42 — Formulation of expression for pulse train envelope in terms of impulse characteristic and amplitudes of transmitted pulses.

more discrete values as in pulse code modulation systems. In pulse position modulation, $A_n = 0$ except at the instants pulses of a given amplitude are transmitted, and n may not necessarily be an integer. In pulse duration modulation, $A_n = 1$ over the intervals $n\tau$ of varying duration in which pulses are transmitted, and zero otherwise. Equation (12.02) is thus a general formulation of the wave shape of a received pulse train, applicable to various pulse modulation methods.

Inserting (2.09) in (12.02) with $R_- + R_+ = R$ and $Q_- - Q_+ = Q$ and taking $\psi_r = 0$ without loss of generality

$$\begin{aligned} W(t_0) &= \sum_{n=-\infty}^{\infty} A_n [\cos \omega_r(t_0 + n\tau)R(t_0 + n\tau) + \sin \omega_r(t_0 + n\tau)Q(t_0 + n\tau)], \\ &= \cos \omega_r t_0 \sum_{n=-\infty}^{\infty} A_n [\cos \omega_r n\tau R(t_0 + n\tau) + \sin \omega_r n\tau Q(t_0 + n\tau)] \\ &\quad + \sin \omega_r t_0 \sum_{n=-\infty}^{\infty} A_n [\cos \omega_r n\tau Q(t_0 + n\tau) - \sin \omega_r n\tau R(t_0 + n\tau)]. \end{aligned} \quad (12.03)$$

The envelope of the wave at the sampling instant $t_0 = 0$ is

$$\bar{W}(0) = (\bar{R}^2 + \bar{Q}^2)^{1/2}, \quad (12.04)$$

$$\bar{R} = \sum_{n=-\infty}^{\infty} A_n [\cos \omega_r n\tau R(n\tau) + \sin \omega_r n\tau Q(n\tau)], \quad (12.05)$$

$$\bar{Q} = \sum_{n=-\infty}^{\infty} A_n [\cos \omega_r n\tau Q(n\tau) - \sin \omega_r n\tau R(n\tau)].$$

For the particular case of a low-pass system

$$Q = 0, \text{ and } \omega_r = 0,$$

so that

$$W(0) = \sum_{n=-\infty}^{\infty} A_n P(n\tau). \quad (12.06)$$

A band-pass characteristic can be obtained with the aid of band-pass filters at the transmitting or receiving ends, or at both ends of a system, and the equations developed previously for the impulse characteristic tacitly assumed such an arrangement. Equivalent performance can, however, also be secured by methods which are usually employed in practice, and to which the equations also apply. Impulses can thus be applied to a low-pass pulse shaping network or filter, and the output used to modulate a carrier. There will then be a symmetrical distribution of

sidebands with respect to the carrier, equivalent to a band-pass characteristic, with the spectrum of the sideband frequencies determined by the characteristic of the low-pass filter. The equivalent of an asymmetrical band-pass characteristic can be obtained by suppressing part of the upper or lower sideband with the aid of filters.

Although the mathematical formulation with both methods is essentially the same when ω_r is identified with the carrier frequency ω_c , with impulse excitation the phase of ω_r is fixed in relation to the envelope but is independent of it with carrier modulation. By proper choice of the pulse interval τ in (12.03), such that $\cos \omega_r(t_0 + n\tau) = \cos \omega_r t_0$ or $\omega_r \tau = 2\pi m$, $m = 0, 1, 2 \dots$ it is possible with impulse excitation to obtain the same relation between the reference or carrier frequency as when the output of a low-pass filter is used to modulate a carrier. In the above case the pulses are transmitted at intervals $\tau = m/f_r = m/f_c$, corresponding to multiples of the duration of a carrier cycle. Since the duration of a carrier cycle is ordinarily small in relation to the pulse interval, there is essentially no important difference in the rate at which pulses can be transmitted with the above two methods. However, with band-pass filters the exact relationship of pulse intervals to the carrier frequency may be difficult to maintain with simple instrumentation, while this is no problem with carrier modulation. For this reason, and since the performance is otherwise equivalent, only the basic relationships with carrier modulation will be discussed further.

Assuming that $\cos \omega_r(t_0 + n\tau) = \cos \omega_r t_0$, as discussed above, equation (12.03) becomes

$W(t_0)$

$$= \cos \omega_r t_0 \sum_{-\infty}^{\infty} A_n R(t_0 + n\tau) + \sin \omega_r t_0 \sum_{-\infty}^{\infty} A_n Q(t_0 + n\tau). \quad (12.07)$$

The envelope at the sampling point is accordingly

$$\bar{W}(0) = \left(\left[\sum_{-\infty}^{\infty} A_n R(n\tau) \right]^2 + \left[\sum_{-\infty}^{\infty} A_n Q(n\tau) \right]^2 \right)^{1/2}. \quad (12.08)$$

In ideal transmission systems there would be no pulse overlaps or intersymbol interference, and the amplitude of the pulse train at the sampling instant would be

$$\bar{W}(0) = A_0 [R^2(0) + Q^2(0)]^{1/2}. \quad (12.09)$$

This condition could be realized with sufficient pulse spacing. However, the objective in the design of efficient pulse systems is to determine the minimum pulse spacing consistent with tolerable intersymbol inter-

ference and thus the maximum transmission capacity or optimum performance in other respects for a given bandwidth. In the following sections this problem is discussed further.

13. TRANSMISSION LIMITATIONS IN SYMMETRICAL SYSTEMS

In a symmetrical system the amplitude characteristic has even symmetry and the phase characteristic odd symmetry with respect to a properly chosen frequency. A low-pass transmission system is thus symmetrical with respect to zero frequency, when the negative frequency range is included. A double sideband system is symmetrical if the amplitude characteristic has even and the phase characteristic odd symmetry with respect to the mid-band frequency.

Equation (12.06) applying to a low-pass system or baseband transmission may be written

$$W(0) = A_0 P(0) + \sum_{n=1}^{\infty} [A_n P(n\tau) + A_{-n} P(-n\tau)]. \quad (13.01)$$

Let it be assumed that pulses of varying but discrete amplitudes are transmitted, with a maximum peak amplitude equal to $A_{\max}$ and a minimum peak amplitude $A_{\min}$. If q pulse amplitudes are employed, the difference between peak amplitudes is then $(A_{\max} - A_{\min})/(q - 1)$. Let P^+ designate positive values of $P(n\tau)$ and P^- the absolute value of negative amplitudes.

The maximum value of $W(0)$ when a pulse of amplitude A_0 is transmitted at the sampling point $n = 0$ is then

$$\begin{aligned} W_{\max} = A_0 P(0) + \sum_{n=1}^{\infty} A_{\max} [P^+(n\tau) + P^+(-n\tau)] \\ - \sum_{n=1}^{\infty} A_{\min} [P^-(n\tau) + P^+(-n\tau)] \end{aligned} \quad (13.02)$$

The minimum amplitude of $W(0)$ when a pulse of the next higher amplitude $A_0 + (A_{\max} - A_{\min})/(q - 1)$ is transmitted becomes

$$\begin{aligned} W_{\min} = \left(A_0 + \frac{A_{\max} - A_{\min}}{q - 1} \right) P(0) \\ - \sum_{n=1}^{\infty} A_{\max} [P^-(n\tau) + P^+(-n\tau)] \\ + \sum_{n=1}^{\infty} A_{\min} [P^+(n\tau) + P^+(-n\tau)]. \end{aligned} \quad (13.03)$$

To permit distinction between the two pulse peaks it is necessary that

$W_{\min}$ be greater than $W_{\max}$. The difference $M = W_{\min} - W_{\max}$, which represents the margin for distinction between pulse amplitudes, becomes

$$M = \frac{A_{\max} - A_{\min}}{q - 1} P(0) - (A_{\max} - A_{\min}) \sum_{n=1}^{\infty} [P^{+}(n\tau) + P^{-}(n\tau) + P^{+}(-n\tau) + P^{-}(-n\tau)], \quad (13.04)$$

or:

$$M = (A_{\max} - A_{\min}) \left[\frac{P(0)}{q - 1} - \sum_{n=1}^{\infty} |P(n\tau)| + |P(-n\tau)| \right], \quad (13.05)$$

where $|P(\pm n\tau)|$ designates the absolute values of the impulse characteristic.

Equation (13.05) shows that for a given value of q the margin depends on the maximum pulse excursion $A_{\max} - A_{\min}$ and is thus the same with $A_{\max} = 1$ and $A_{\min} = 0$ as with $A_{\max} = 0.5$ and $A_{\min} = -0.5$. As an example, equation (13.05) shows that with two pulse amplitudes, $q = 2$, it is possible to distinguish between pulses and spaces, or between positive and negative pulses, if the sum of the absolute values of the impulse characteristic at all the sampling points, excluding 0, is less than the amplitude $P(0)$ of the impulse at sampling point 0.

The maximum margin against errors is obtained without pulse overlaps, i.e. when the summation term in (13.05) is zero, and is

$$M_{\max} = (A_{\max} - A_{\min}) \frac{P(0)}{q - 1}. \quad (13.06)$$

The ratio of the margin M as given by (13.05) to the maximum margin becomes:

$$M/M_{\max} = 1 - \frac{q - 1}{P(0)} \sum_{n=1}^{\infty} |P(n\tau)| + |P(-n\tau)|. \quad (13.07)$$

This equation may be employed to determine the maximum possible pulsing rate for a given impulse characteristic and number of pulse amplitudes, obtained when $M/M_{\max} = 0$, or to determine the margin for a given pulse transmission rate. An example of the latter application is illustrated in Fig. 43, which shows the margin $M/M_{\max}$ in per cent, obtained when (13.07) with $q = 2$ is applied to the curves shown in Fig. 23 for various degrees of delay distortion. The pulse interval is taken as $\tau = 1/2f_1 = 1/f_{\max}$, where f_1 is the frequency at the 6 db down

point on the amplitude characteristic and $f_{\max} = 2f_1$. Under this condition there is no intersymbol interference in the absence of phase distortion.

The above equations apply to peak intersymbol interference, obtained by taking the maximum positive and negative values of the summation term in (13.01). As discussed in previous sections, certain types of transmission system imperfections give rise to pulse distortion extending over long time intervals, such as fine structure deviations over the transmission band, a low-frequency cut-off and pronounced band-edge phase deviations. Evaluation of peak intersymbol interference is then rather difficult, and a more convenient approximate method is to evaluate rms intersymbol interference, which can be related to rms deviation in the transmission frequency characteristic by methods discussed previously. Peak intersymbol interference may then be estimated by applying a peak factor between 3 and 4, depending on the type of transmission distortion.

If $P_0(n\tau)$ designates an ideal impulse characteristic, which is zero for $n = \pm 1, \pm 2$ etc., the deviation from the ideal envelope of a pulse

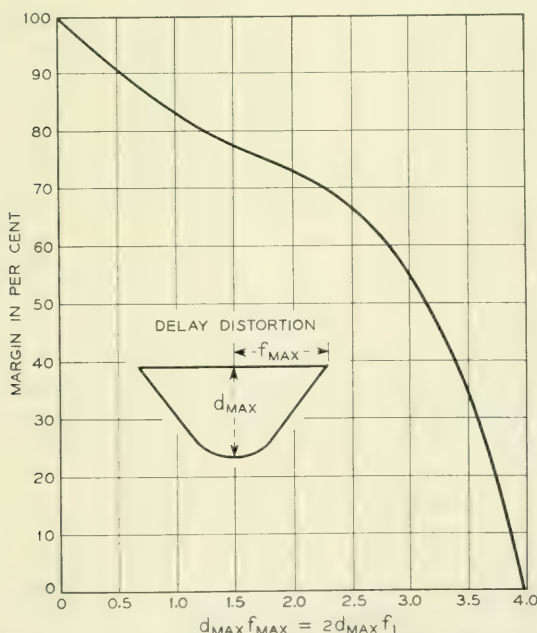


Fig. 43 — Margin against excessive peak interference in systems employing two pulse amplitudes with intervals between pulses $\tau = \tau_1 = 1/2f_1 = 1/f_{\max}$, for impulse transmission characteristic as shown in Fig. 23.

train may be written

$$\Delta W(0) = W(0) - W_0(0) = \sum_{n=-\infty}^{\infty} A_n [P(n\tau) - P_0(n\tau)]. \quad (13.08)$$

The rms deviation becomes, with $\Delta P(n\tau) = P(n\tau) - P_0(n\tau)$

$$\underline{\Delta W}(0) = \underline{A} \left(\sum_{n=-\infty}^{\infty} [\Delta P(n\tau)]^2 \right)^{1/2}, \quad (13.09)$$

$$\cong \underline{A} \left(\frac{1}{\tau} \int_{-\infty}^{\infty} [\Delta P(t)]^2 dt \right)^{1/2}, \quad (13.10)$$

$$= \underline{A} P(0) \underline{U}. \quad (13.11)$$

$\underline{A}$ is the rms amplitude of the transmitted pulses and $\underline{U}$ the rms intersymbol interference referred to unit amplitude of the received pulses. Expressions for $\underline{U}$ applying to fine structure imperfections in the transmission frequency characteristic were given in Section 8, for a low-frequency cut-off in Section 9 and for band-edge phase deviations in Section 10.

For balanced pulse systems employing positive and negative pulses, rms intersymbol interference in the positive and negative directions will be equal. For such systems the maximum value of the summation in (13.02) becomes $k\underline{W}(0)$ and in (13.03) $-k\underline{W}(0)$, where k is the peak factor. Equation (13.04) is then replaced by

$$\begin{aligned} M &= \frac{A_{\max} - A_{\min}}{q - 1} P(0) - 2k\underline{A}P(0)\underline{U}, \\ &= 2A_{\max}P(0) \left[\frac{1}{q - 1} - k\underline{U}(\underline{A}/A_{\max}) \right], \end{aligned} \quad (13.12)$$

when $A_{\min} = -A_{\max}$.

In a balanced pulse system employing q pulse amplitudes, i.e., $q/2$ positive and $q/2$ negative amplitudes, with equal steps $2A_{\max}/(q - 1)$ between pulse amplitudes, the following relation applies if all amplitudes have equal probability.

$$\underline{A}/A_{\max} = \left[\frac{q + 1}{3(q - 1)} \right]^{1/2}. \quad (13.13)$$

Hence,

$$M = \frac{2A_{\max}P(0)}{q - 1} \left[1 - k \left(\frac{q^2 - 1}{3} \right)^{1/2} \underline{U} \right]. \quad (13.14)$$

As mentioned before, the factor k may be as high as 4, in which case the

maximum tolerable rms intersymbol interference $\underline{U}$ referred to unit peak amplitude of the received pulses becomes for $M = 0$:

$$\begin{array}{ccc} q = \underline{2} & \underline{4} & \underline{8} \\ \underline{U} = 0.25 & 0.112 & 0.054 \end{array}$$

In (13.14) and in the above table, $\underline{U}$ is the maximum tolerable rms intersymbol interference from all sources, such as fine structure imperfections over the transmission band, band-edge phase distortion and a low-frequency cut-off. Interference from these various sources may be combined on a root-sum-square basis.

In the above evaluation of rms intersymbol interference a balanced pulse system was assumed. An unbalanced system can be obtained by superposing on a balanced system an infinite sequence of pulses of equal amplitude and polarity at uniform intervals as indicated in Fig. 44. This superposed system will give rise to a fixed intersymbol interference or displacement of the pulse train, which does not alter the margin for distinction between pulse amplitudes and which can be corrected by a fixed bias at the receiving end if necessary. For this reason, in the case

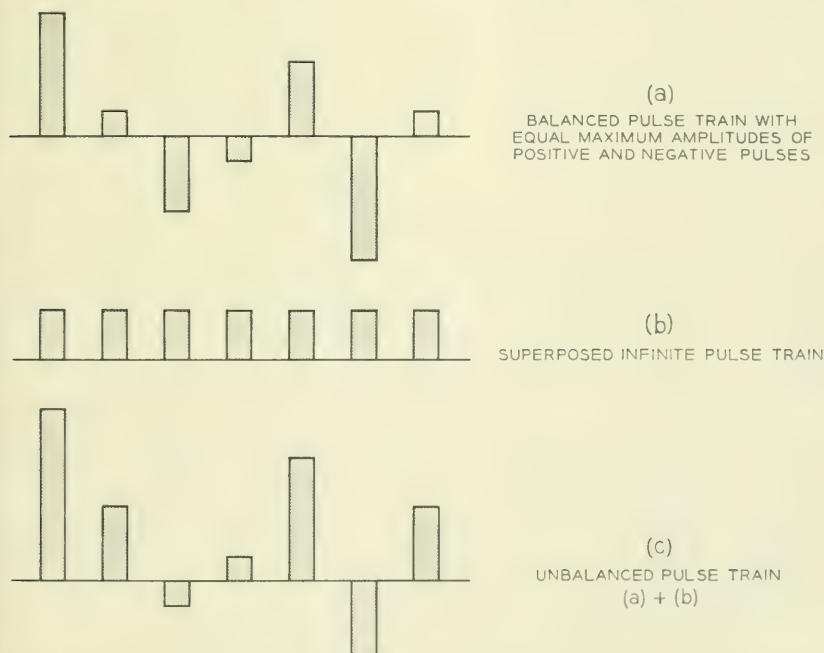


Fig. 44 Derivation of an unbalanced from a balanced pulse train by superposition of an infinite train of pulses of equal amplitude.

of an unbalanced system, only the balanced component need to be considered in evaluating rms intersymbol interference, which will thus be the same whether or not the system is balanced. As shown previously, peak intersymbol interference, or the margin for distinction between pulse amplitudes, depends only on the peak to peak pulse excursion and is thus the same for unbalanced as for balanced systems. It may be noted here that for a balanced system the transmitted power is a minimum for a given margin in pulse reception, as is the interference in other systems that may be caused by the transmitted pulses.

For a symmetrical band-pass system, rather than a low-pass system as discussed above, $Q(n\tau) = 0$ in (12.08). The envelope of the pulse train then becomes

$$\bar{W}(0) = \sum_{n=-\infty}^{\infty} A_n R(n\tau), \quad (13.15)$$

where $R(n\tau) = R_-(n\tau) + R_+(n\tau) = 2R_+(n\tau)$, with R_- and R_+ given by (2.10).

Since (13.15) is of the same form as (13.01), the relationships established above for low-pass systems also apply to symmetrical band-pass systems, with $R(n\tau)$ replacing $P(n\tau)$. $R(n\tau)$ will have the same shape as $P(n\tau)$, but will be greater by a factor 2, which will appear as a multiplier in the various expressions and hence not alter the requirements on tolerable pulse distortion or intersymbol interference.

14. TRANSMISSION LIMITATIONS IN ASYMMETRICAL SIDEBAND SYSTEMS

The formulation of transmission limitations imposed by pulse distortion in asymmetrical sideband systems is complicated by the presence of the quadrature component in the impulse transmission characteristic. Of particular interest are the transmission limitations with vestigial sideband as compared with double sideband transmission, assuming the same bandpass characteristic in both cases, a question which has been dealt with in literature for systems with a linear phase characteristic^{4, 11}. Relationships (2.18) and (2.19) facilitate a comparison also for systems with phase distortion, as shown in the following.

If the envelope of the impulse characteristic with double sideband transmission is $\bar{P}(t)$, the in-phase and quadrature components with vestigial sideband transmission are given by (2.19), with $\omega_y = \omega_s$ or

$$\begin{aligned} R &= R_- + R_+ = \cos(\omega_s t - \psi_s) \bar{P}(t), \\ Q &= Q_- - Q_+ = \sin(\omega_s t - \psi_s) \bar{P}(t). \end{aligned} \quad (14.01)$$

If t is so chosen that $\omega_s t - \psi_s = 0$, and the time with respect to this value of t is designated t_0 , then

$$\begin{aligned} R(t_0) &= \cos \omega_s t_0 \bar{P}(t_0), \\ Q(t_0) &= \sin \omega_s t_0 \bar{P}(t_0). \end{aligned} \quad (14.02)$$

An application of this method to the impulse characteristic shown in Fig. 23 for $b = 15$ radians is illustrated in Fig. 45.

In order to compare vestigial with double sideband transmission, it suffices to evaluate the in-phase and quadrature components at the sampling instants. With $\tau = \pi/2\omega_s = 1/4f_s$, the in-phase and quadrature components at times $m\tau$, for $m = 0 \pm 1, \pm 2$, etc., will be as illustrated in Fig. 46.

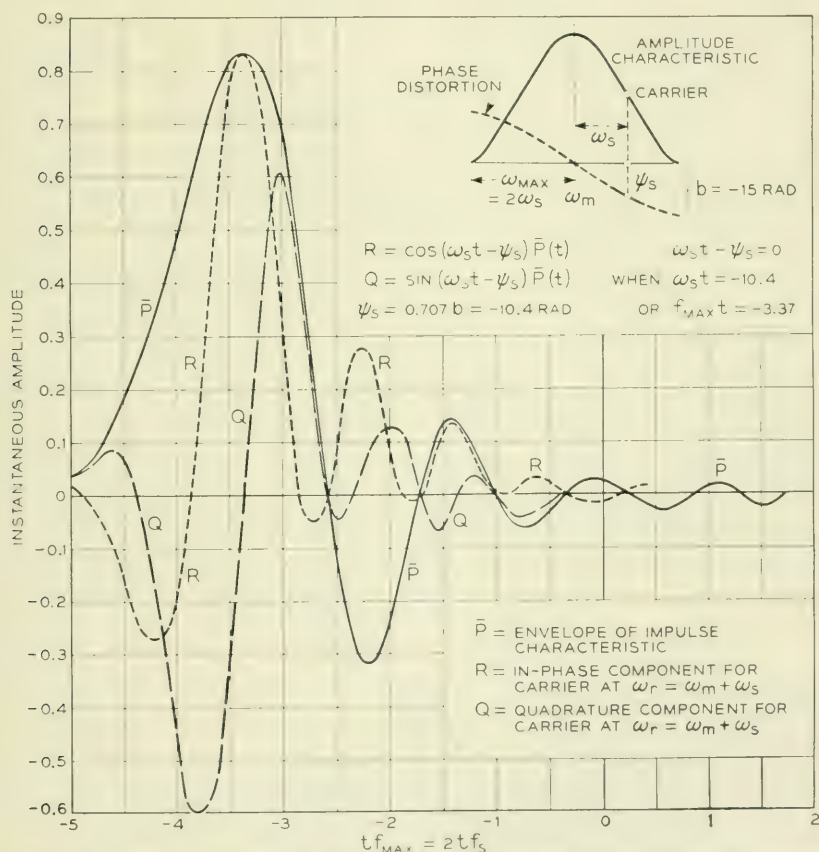


Fig. 45 — Determination of in-phase and quadrature components of impulse characteristic for vestigial side-band transmission.

characteristic having a linear phase shift, there would be no intersymbol interference with either method for the above rate of pulse transmission.

Assume next that the pulse transmission rate is doubled and that the quadrature component is eliminated. This is possible if the carrier frequency is transmitted and is derived at the receiving end with the aid of filters and applied in proper phase to a product demodulator, a method known as homodyne detection. At the points $m = 1, 3, 5$, etc., there would then be no quadrature components and no in-phase components. The sum of the absolute values of the in-phase components at the other sampling points, $m = 2, 4$, etc., would be the same as with double sideband transmission. It follows that the transmission capacity (pulsing rate) can be doubled by vestigial sideband transmission if the quadrature component is eliminated by homodyne detection, for the same margin against excessive intersymbol interference as with double sideband transmission.

An increase in transmission capacity can be realized with vestigial sideband transmission without elimination of the quadrature component by homodyne detection, although a two-fold increase is then possible only if the phase characteristic is linear, as discussed below. Vestigial sideband transmission can be employed without transmission of the carrier, or with a fixed level of carrier in the absence of pulses and a higher level in the presence of pulses. The latter method is equivalent to the transmission of two or more pulse amplitudes, with the minimum amplitude greater than zero. With this method the effect of the quadrature component on the envelope of a pulse train can be reduced, and even eliminated provided the phase characteristic is linear. In the following, vestigial sideband transmission with two pulse amplitudes at twice the double sideband pulsing rate is discussed, for the case in which the minimum pulse amplitude is finite rather than zero.

With pulses transmitted at twice the double sideband rate, i.e., with the interval between pulses equal to $\tau = \pi/2\omega_s$, equation (12.08) for the envelope becomes in view of (14.02)

$\bar{W}(0)$

$$= \left(\left[\sum_{-\infty}^{\infty} A_n \cos \omega_s n \tau \bar{P}(n\tau) \right]^2 + \left[\sum_{-\infty}^{\infty} A_n \sin \omega_s n \tau \bar{P}(n\tau) \right]^2 \right)^{1/2}. \quad (14.03)$$

At the even sampling points, i.e., $n = 0, 2, 4 \dots$, $\cos \omega_s n \tau = \pm 1$ and the in-phase components may be written

$$R(\pm 2m\tau) = \pm \bar{P}(\pm 2m\tau), \quad m = 0, 1, 2 \dots$$

At the odd sampling points, i.e., $n = 1, 3, 5 \dots$, $\sin \omega_s n \tau = \pm 1$ and

the quadrature components may be written

$$Q[\pm(2m-1)\tau] = \pm \bar{P}[\pm(2m-1)\tau], \quad m = 1, 2, 3$$

Let

$$\begin{aligned} \sum R^+ &= \sum_{m=1}^{\infty} [R^+(2m\tau) + R^+(-2m\tau)], \\ \sum R^- &= \sum_{m=1}^{\infty} [R^-(2m\tau) + R^-(-2m\tau)], \\ \sum Q^+ &= \sum_{m=1}^{\infty} Q^+[(2m-1)\tau] + Q^+[-(2m-1)\tau], \\ \sum Q^- &= \sum_{m=1}^{\infty} Q^-[(2m-1)\tau] + Q^-[-(2m-1)\tau], \end{aligned} \quad (14.04)$$

where R^+ , Q^+ designate positive values and R^- , Q^- the absolute values of negative amplitudes of the in-phase and quadrature components.

Let it be assumed that two pulse amplitudes are employed, $A_{\min}$ and $A_{\max}$. When the minimum amplitude is transmitted, the maximum value of the envelope is obtained by considering the maximum positive overlaps of the in-phase components in conjunction with the maximum value of the quadrature component. The value thus obtained is

$$\begin{aligned} \bar{W}_{\max} &= [(A_{\min} R(0) + A_{\max} \sum R^+)^2 \\ &\quad + (A_{\max} \sum Q^- - A_{\min} \sum Q^+)^2]^{1/2} \end{aligned} \quad (14.05)$$

It is assumed that $\sum Q^- > \sum Q^+$, otherwise Q^- and Q^+ would be interchanged in the last term.

When the maximum amplitude is transmitted, the minimum value of the envelope is obtained by considering the maximum negative overlaps of the in-phase components, in conjunction with the minimum value of the quadrature component, which gives

$$\begin{aligned} \bar{W}_{\min} &= [(A_{\max} R(0) - A_{\max} \sum R^-)^2 + A_{\min}^2 (\sum Q^- \\ &\quad - \sum Q^+)^2]^{1/2}. \end{aligned} \quad (14.06)$$

The margin for distinction between $A_{\min}$ and $A_{\max}$ is $M = W_{\min} - W_{\max}$ and becomes

$$\begin{aligned} M &= A_{\max} [(R(0) - \sum R^-)^2 + \mu^2 (\sum Q^- - \sum Q^+)^2]^{1/2} \\ &\quad - A_{\max} [(\mu R(0) + \sum R^+)^2 + (\sum Q^- - \mu \sum Q^+)^2]^{1/2}, \end{aligned} \quad (14.07)$$

where

$$\mu = A_{\min}/A_{\max}.$$

The margin for a unit difference $A_{\max} - A_{\min}$, i.e. $M_1 = M/(A_{\max} - A_{\min})$ becomes:

$$M_1 = \frac{1}{1 - \mu} ([R(0) - \sum R^-]^2 + \mu^2 [\sum Q^- - \sum Q^+]^2)^{1/2} \quad (14.08)$$

$$- [(\mu R(0) + \sum R^+)^2 + (\sum Q^- - \mu \sum Q^+)^2]^{1/2}.$$

The special case of an ideal transmission characteristic as shown in Fig. 47 will be considered first. In this case

$$\begin{aligned} R(0) &= 1 & R(2\tau) &= 0 & R(-2\tau) &= 0 \\ R(4\tau) &= 0 & R(-4\tau) &= 0 \\ Q(\tau) &= 0.5 & Q(-\tau) &= -0.5 \\ Q(3\tau) &= 0 & Q(-3\tau) &= 0 \end{aligned}$$

so that:

$$\begin{aligned} \sum R^+ &= 0 & \sum R^- &= 0 \\ \sum Q^+ &= 0.5 & \sum Q^- &= 0.5 \end{aligned}$$

Equation (14.08) in this case simplifies to

$$M_1 = \frac{1}{1 - \mu} \left(1 - \left[\mu^2 + \frac{1}{4} (1 - \mu)^2 \right]^{1/2} \right). \quad (14.09)$$

For various values of $\mu = A_{\min}/A_{\max}$ the margin for unit difference in

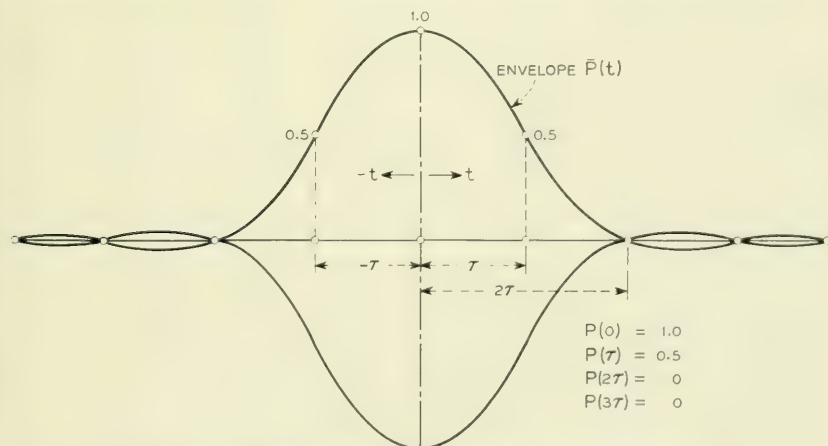


Fig. 47 — Envelope of idealized impulse characteristic.

pulse amplitudes becomes:

μ	0	0.2	0.3	0.5	0.8	0.9	1.0
M_1	0.50	0.69	0.77	0.87	0.96	0.998	1.0

Thus, for an ideal impulse characteristic as assumed above, the quadrature component gives rise to 50 per cent maximum intersymbol with $\mu = 0$, and to negligible intersymbol interference when $\mu \geq 0.8$ or greater. By way of comparison, the margin would be zero with double sideband transmission at the rate assumed above, i.e., twice the normal double sideband rate. This follows from (13.07) when it is considered that $P(\pm\tau) = \frac{1}{2} P(0)$, $P(\pm 2\tau) = 0$, so that the sum of the absolute values of the impulse characteristic at the sampling points is equal to $P(0)$ and thus $M/M_{\max} = 0$.

Elimination of the effect of the quadrature component by the above method is contingent on a symmetrical impulse characteristic, i.e., $\bar{P}(n\tau) = \bar{P}(-n\tau)$, a condition which can be realized only with a linear phase shift. Furthermore, the in-phase components must vanish at the sampling points, which entails an ideal amplitude characteristic. In the presence of phase distortion the effect of the quadrature component cannot be eliminated but may be reduced by proper choice of the ratio μ , as discussed below for a transmission characteristic with moderate phase distortion.

As an example consider an impulse characteristic as shown in Fig. 23 for $b = 5$ radians. The in-phase and quadrature components at the various sampling points are in this case

$$\begin{aligned}
 R(0) &= 0.97 & R(-2\tau) &= -0.09 & R(2\tau) &= 0.13 \\
 R(-4\tau) &\cong 0 & R(4\tau) &\cong 0 \\
 Q(-\tau) &= -0.54 & Q(\tau) &= 0.44 \\
 Q(-3\tau) &\cong 0 & Q(3\tau) &= -0.03 \\
 Q(-5\tau) &\cong 0 & Q(5\tau) &\cong 0
 \end{aligned}$$

Hence

$$\sum R^+ = 0.13 \quad \sum R^- = 0.09 \quad \sum Q^+ = 0.44 \quad \sum Q^- = 0.57$$

Equation (14.08) in this case becomes

$$\begin{aligned}
 M_1 = \frac{1}{1 - \mu} & \left((0.88^2 + 0.13^2 \mu^2)^{1/2} - [(0.97\mu + 0.13)^2 \right. \\
 & \left. + (0.57 - 0.44\mu)^2]^{1/2} \right).
 \end{aligned}$$

For various values of $\mu = A_{\min}/A_{\max}$ the margin for unit difference in pulse amplitudes becomes

μ	0	0.2	0.3	0.4	0.5	0.6	0.7	0.75
M_1	0.30	0.375	0.40	0.375	0.34	0.25	0.13	0

The optimum condition is thus in the above particular case obtained with $\mu \cong 0.3$, with a comparatively small variation in transmission performance for any value of μ between 0 and 0.5.

In the above discussion of vestigial sideband transmission, modulation of a carrier was assumed, with elimination of one sideband except for the wanted vestige. The equivalent performance can be secured by application of impulses to a band-pass transmission characteristic with the proper interval between pulses in relation to the midband frequency, as discussed below:

When equation (12.03) is written with respect to the midband frequency, $\omega_r = \omega_m$, and a symmetrical amplitude characteristic is assumed so that $Q = 0$, the following relation obtained.

$$\begin{aligned}
 W(t_0) = & \cos \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \cos \omega_m n\tau R(t_0 + n\tau) \\
 & - \sin \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \sin \omega_m n\tau R(t_0 + n\tau),
 \end{aligned}
 \tag{14.10}$$

in which R may be replaced by $\bar{P}$, the envelope of the impulse characteristic.

Let it be assumed that τ is so chosen that $\cos \omega_m n\tau = \cos n\pi/2$ in which case $\sin \omega_m n\tau = \sin n\pi/2$. The above equation then becomes

$$\begin{aligned}
 W(t_0) = & \cos \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \bar{P}(t_0 + n\tau) \cos n\pi/2 \\
 & - \sin \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \bar{P}(t_0 + n\tau) \sin n\pi/2.
 \end{aligned}
 \tag{14.11}$$

The in-phase and quadrature components of the envelope at the sampling instant $t_0 = 0$ are accordingly

$$\begin{aligned}
 \bar{R}(0) &= \sum_{n=-\infty}^{\infty} A_n \bar{P}(n\tau) \cos n\pi/2, \\
 \bar{Q}(0) &= \sum_{n=-\infty}^{\infty} A_n \bar{P}(n\tau) \sin n\pi/2.
 \end{aligned}
 \tag{14.12}$$

Pulses in even positions, i.e., A_0 , A_2 , A_4 , etc., will thus contribute an in-phase but no quadrature component while pulses in odd positions

A_1, A_3, A_5 , etc., will contribute a quadrature but no in-phase component. It will be recognized from Fig. 42 that this is the same condition as encountered in vestigial sideband transmission with pulses in the latter case transmitted at intervals $\tau = \pi/2\omega_s = 1/4 f_s$.

To realize the above condition with pulses applied to a band-pass filter, it is necessary that in (14.10)

$$\omega_m n \tau = n\pi(\frac{1}{2} + N), \quad (14.13)$$

where N is an integer, or that

$$\tau = \frac{\pi(1 + 2N)}{2\omega_m} = \frac{1 + 2N}{4f_m}. \quad (14.14)$$

The interval between pulses must thus be an integral number of half-cycles plus one quarter cycle of the midband frequency f_m , as illustrated for a particular case in Fig. 48. When f_m is large in relation to the sideband frequency this condition can be achieved with substantially the same pulse spacing as with vestigial sideband transmission. To secure exactly the same rate of pulse transmission it is necessary that

$$\tau = 1/4 f_s,$$

which, in conjunction with (14.14), gives

$$N = \frac{1}{2} (f_m/f_s - 1). \quad (14.15)$$

Thus, if $f_m = 5f_s$, $N = 2$ and the interval τ between pulses as obtained from (14.14) is 1.25 cycles of f_m . If $f_m = 10f_s$, $N = 4.5$ and it is not possible to have exactly the same pulsing rate as with vestigial sideband transmission, since N must be an integer. It is then necessary to take $N = 4$ or 5. With $N = 4$ equation (14.14) gives $\tau = 9/40f_s$ and with $N = 5$, $\tau = 11/40f_s$. This compares with $\tau = 1/4f_s = 10/40f_s$ with vestigial sideband transmission, so that there is a minor difference in pulse intervals with the two methods.

15. DOUBLE VERSUS VESTIGIAL SIDEBAND SYSTEMS

From the preceding discussion it follows that, for the same bandwidth and margin against interference from characteristic distortion, a two-fold increase in transmission capacity can be approached with vestigial over double sideband transmission. This assumes that the carrier is transmitted at the proper level and that the phase characteristic is linear, or that otherwise homodyne detection is used to cancel the effect of the quadrature component.

For the same bandwidth, the same transmission capacity can be realized with a double sideband system employing four pulse amplitudes as with a vestigial sideband system with two pulse amplitudes. However, the latter type of system will have a greater tolerance to interference from characteristic distortion than the former. This follows when it is considered that in a quaternary system the maximum tolerable interference is $\frac{1}{6}$ the maximum pulse amplitude, as compared to $\frac{1}{2}$ the maximum pulse amplitude in a binary system. With $\mu = A_{\min}/A_{\max} = 0$, the quadrature component reduces the margin by a factor of 0.5, so that the maximum tolerable interference in relation to the maximum pulse

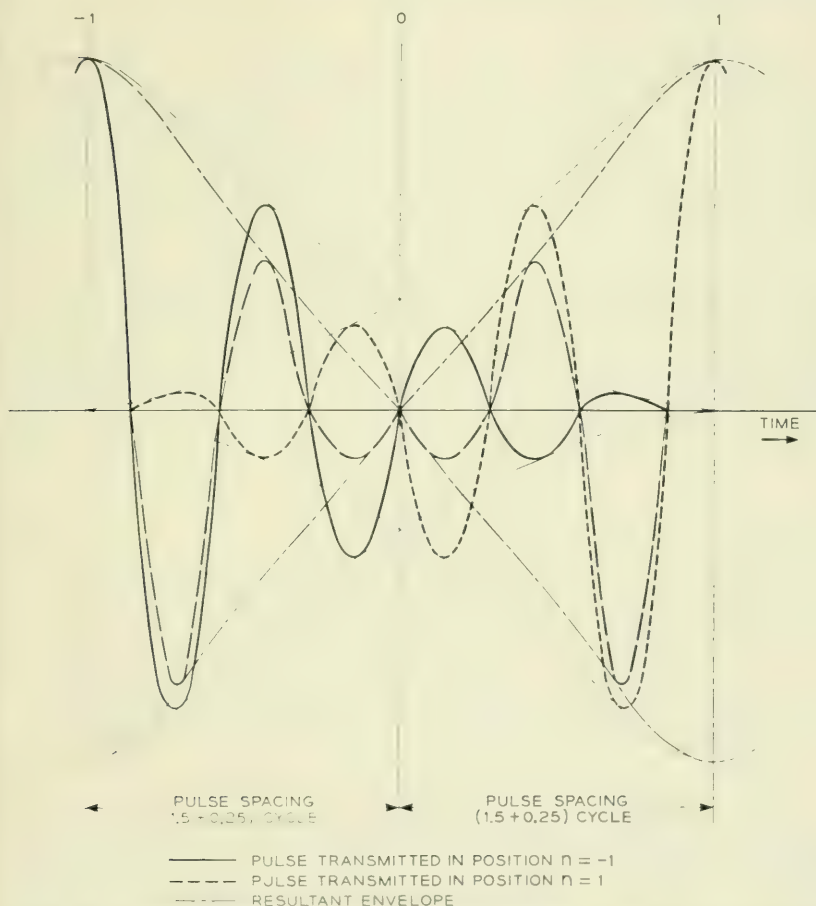


Fig. 48 — Impulse excitation of band-pass system with pulse spacing selected to provide equivalent of vestigial side-band transmission.

amplitude is $\frac{1}{4}$ as compared to $\frac{1}{6}$ for a quaternary system. If the phase characteristic is linear and the carrier is transmitted at the optimum level, or if homodyne detection is used, the effect of the quadrature component is cancelled. The maximum tolerable interference is then $\frac{1}{2}$ as compared with $\frac{1}{6}$ for a quaternary double sideband system.

In the presence of phase distortion, a substantial advantage can also be realized with a binary vestigial system, which can be illustrated by considering the example in Section 14. For the optimum condition $\mu = 0.4$, the margin is reduced by a factor 0.4 and is thus 0.2. For a quaternary double sideband system the factor by which the margin is reduced is given by (13.07), with $q = 4$ and with

$$\frac{1}{P(0)} \sum_{n=1}^{\infty} |P(n\tau)| + |P(-n\tau)| = \frac{1}{R(0)} [\sum R^+ + \sum R^-],$$

where $R(0) = 0.97$, $\sum R^+ = 0.13$ and $\sum R^- = 0.09$, as in the example in Section 14. The reduction in margin thus obtained is $M/M_{\max} = 0.32$. Hence the maximum tolerable interference for a quaternary double sideband system is $0.32/6 = 0.053$ as compared with 0.20 for a binary vestigial sideband system under the optimum condition $\mu = 0.4$.

For the same transmission capacity and same number of pulse amplitudes, a substantial transmission advantage may be realized with vestigial over double sideband transmission in circuits with pronounced phase distortion, owing to the circumstance that a two-fold reduction in bandwidth with vestigial sideband transmission may afford a sub-

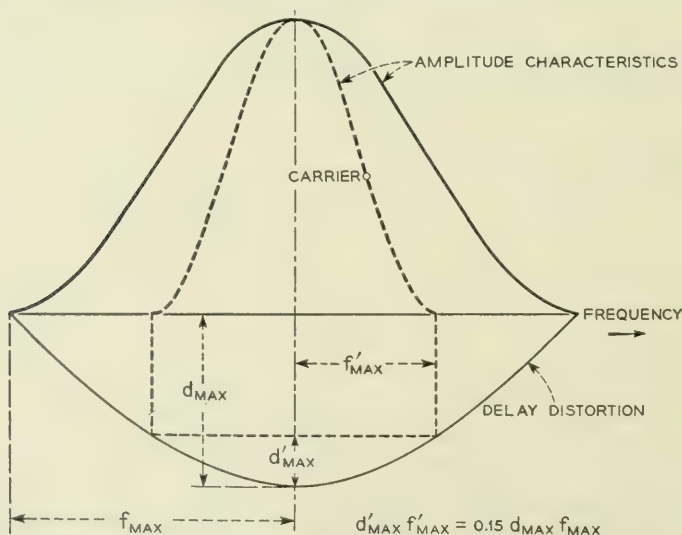


Fig. 49 — Comparison of double and vestigial side-band transmission in the presence of delay distortion.

stantial reduction in delay distortion over the transmission band. This is illustrated in Fig. 49, where a cosine variation in transmission delay is assumed. With a two-fold reduction in bandwidth, the product $d'_{\max}f'_{\max}$ for vestigial sideband transmission is about 15 per cent of the product $d_{\max}f_{\max}$ for double sideband transmission. Thus, with $d_{\max}f_{\max} = 8.3$, $d'_{\max}f'_{\max} = 1.25$, corresponding to $b = 5$ radians, as assumed in the example in Section 14. Vestigial sideband transmission is in this case feasible with an adequate margin, about 40 per cent of the maximum margin in the absence of phase distortion. Double sideband transmission would not be possible, as is evident from Fig. 43, since it would be necessary to have $d_{\max}f_{\max}$ less than 4, as compared with 8.3 in the above case.

The above discussion of vestigial vs double sideband transmission pertains to the effects of characteristic distortion rather than noise, and the relative complexity of terminal equipment was disregarded. Because of the simpler terminal equipment with double sideband transmission, this method is ordinarily used where bandwidth is not a primary consideration, as for example in providing a number of telegraph¹ channels over a voice frequency circuit.

16. LIMITATION ON CHANNEL CAPACITY BY CHARACTERISTIC DISTORTION

For an idealized channel of bandwidth f_1 with a transmission-frequency characteristic as shown in Fig. 7, the transmission capacity in bits per second for a signal of average power P in the presence of random noise of average power N can with sufficiently complicated encoding methods approach the limiting value given by Shannon:¹²

$$C = f_1 \log_2 (1 + P/N). \quad (16.01)$$

The above expression also applies to certain other idealized channels with a linear phase characteristic, when f_1 is defined as in Fig. 10. In all of these cases the integral of the area under the amplitude characteristic, or the equivalent bandwidth, is f_1 .

By way of comparison, for pulse code modulation systems the channel capacity is of the same basic form as (16.01), namely¹³:

$$C = f_1 \log_2 \left(1 + \frac{P}{KN} \right), \quad (16.02)$$

where $K \cong 8$. Thus about an 8-fold increase in signal power is required to attain the same channel capacity as with the idealized but impracticable encoding system underlying (16.01).

The above expressions give the limitation on channel capacity imposed by random noise. From the discussion in Sections 13 and 14 it follows that a limitation is placed on channel capacity by characteristic distortion.

tion, in the absence of noise. In idealized communication theory, characteristic distortion has been disregarded in determining channel capacity on the premise that unlike random noise it is predictable and can therefore be corrected, at least in principle. In actual systems, however, complete elimination though possible in principle cannot be accomplished in practice. The resultant limitation on transmission capacity may be as important as that imposed by the maximum signal power that can actually be provided to override noise.

In the following it will be assumed that correction of amplitude and phase deviations is made by equalization, so that the amplitude and phase characteristics are as assumed for an ideal channel, except for small fine structure residual deviations as illustrated in Fig. 30. These small fine structure deviations may be regarded as of random nature in the sense that they differ among channels and cannot be predicted, although for a given system they would remain fixed in the absence of temperature variations or changes in amplifiers with age.

From equation (13.12) it follows that the maximum number of pulse amplitudes or quantizing levels as limited by characteristic distortion is obtained from the relation

$$\frac{1}{q-1} = k\bar{U}\bar{A}/A_{\max}, \quad (16.03)$$

or

$$q = 1 + \frac{1}{k\bar{U}} A_{\max}/\bar{A}. \quad (16.04)$$

In the absence of characteristic distortion, the maximum number of pulse amplitudes as limited by an rms noise amplitude $\underline{A}_n$ or a peak noise amplitude $k\underline{A}_n$ is obtained from the following relation for a balanced pulse system.

$$\frac{A_{\max}}{q-1} = k\underline{A}_n, \quad (16.05)$$

or

$$q = 1 + \frac{\bar{A}}{k\underline{A}_n} A_{\max}/\bar{A}. \quad (16.06)$$

Comparison of (16.04) and (16.06) shows the following equivalence between intersymbol interference and noise from the standpoint of limitation on the permissible number of pulse amplitudes

$$\bar{U} = \underline{A}_n/\bar{A}, \quad (16.07)$$

or

$$\underline{U}^2 = D = N/P. \quad (16.08)$$

This means that random characteristic distortion has the same effect as a random noise power $N = DP$, where D is a distortion factor.

In view of the above equivalence, the channel capacity of a PCM system in the presence of random characteristic distortion, but without noise, as obtained by substitution of (16.08) in (16.02) becomes

$$C = f_1 \log_2 \left(1 + \frac{1}{KD} \right). \quad (16.09)$$

With random interference from both characteristic distortion and noise, the interfering powers add directly, so that for a PCM system

$$C = f_1 \log_2 \left(1 + \frac{1}{K(D + N/P)} \right). \quad (16.10)$$

The equivalence (16.08) was established above on the basis of discrete pulse amplitudes, but it is independent of q and would thus apply also for continuous signals. On this basis it would apply for any method of modulation or of encoding signals and the maximum channel capacity as given by (16.01) would be modified to

$$C = f_1 \log_2 \left(1 + \frac{1}{D + N/P} \right), \quad (16.11)$$

It follows from the above that for any modulation method the tolerable distortion factor is directly related to the average signal-to-noise ratio. Thus two modulation methods which are equivalent from the standpoint of signal-to-noise ratio are also equivalent from the standpoint of tolerable rms distortion, provided faithful reproduction of the transmitted signal is required, as assumed here.

From (8.14) the following relation is obtained between the distortion factor $D = \underline{U}^2$ and small rms deviations $\underline{a}$ (nepers) and $\underline{b}$ (radians) in the amplitude and phase characteristics

$$D = \underline{a}^2 + \underline{b}^2. \quad (16.12)$$

In order that characteristic distortion may be disregarded in comparison with noise, it is necessary that $D \ll N/P$ or

$$\underline{a}^2 + \underline{b}^2 \ll N/P. \quad (16.13)$$

For example, in communication systems employing the same bandwidth as the original signal, such as a pulse amplitude modulation system, a representative signal-to-noise ratio would be about 40 db, or $N/P = 10^{-4}$. In order that characteristic distortion may be disregarded in this case, it would be necessary for both $\underline{a}$ and $\underline{b}$ to be substantially

less than 10^{-2} nepers and radians respectively. This would correspond to an rms gain deviation over the transmission band well below 0.08 db and an rms deviation from a linear phase characteristic well below 0.6 degrees. Since these tolerances are difficult to realize in actual systems, at least for wire circuits, characteristic distortion rather than noise may impose a limitation on channel capacity of systems employing about the same bandwidth as the original signal.

In accordance with (16.01), the bandwidth can in principle be halved without change in channel capacity if the signal-to-noise ratio is squared, i.e., if $N/P = 10^{-8}$ rather than 10^{-4} in the previous example. The tolerable rms amplitude and phase deviations would then be

$$\underline{a}^2 + \underline{b}^2 \ll 10^{-8}.$$

Thus both a and b would have to be substantially smaller than about 10^{-4} , which would preclude a substantial bandwidth saving in practical systems from the standpoint of characteristic distortion, even if it were feasible from the standpoint of signal power required to override noise.

The above considerations apply when faithful reproduction of the transmitted signal is required, as for example in data transmission. In speech transmission considerable distortion can be tolerated, a circumstance which permits appreciable phase distortion in the usual frequency division system without noticeable impairment of intelligibility, but which cannot be taken advantage of in time division pulse systems because of the resultant intersymbol interference. The characteristics of speech sounds also permit a reduction in the bandwidth of the original transmitted signal, by such devices as vocoders or frequency companders, without excessive impairment of intelligibility.

ACKNOWLEDGMENTS

This paper is based on studies in connection with the application of pulse systems to wire circuits, suggested by M. L. Almquist and J. T. Dixon and carried out under their direction. Some of these studies were made on behalf of the Signal Corps, under contract W-36-039-SC-38115. The Signal Corps Engineering Laboratories have consented to the publication of results obtained in these studies. The writer had the benefit in these studies of discussions with C. B. Feldman and W. R. Bennett, and of some of their memoranda on pulse modulation systems. He is also thankful for comments and advice from F. B. Llewellyn and others in connection with preparation of the paper.

REFERENCES

See Part I.

Bell System Technical Papers Not Published in this Journal.

ALLIS, W. P., see ROSE, D. J.

ANDERSON, P. W.,¹ MERRITT, F. R.,¹ REMEIK, J. P.,¹ and YAGER, W. A.¹

Magnetic Resonance in Fe_2O_3 , Phys. Rev., **93**, pp. 717-718, Feb. 15, 1954.

BICKELHAUPT, C. O.²

Fifty Years Ago, Telephony, **146**, pp. 20-22, Feb. 20, 1954.

CLOGSTON, A. M.¹ and HEFFNER, H.¹

Focussing of an Electron Beam by Periodic Fields, J. Appl. Phys., **25**, pp. 436-447, April, 1954.

CONWELL, E. M., see DEBYE, P. P.

DARROW, K. K.¹

Solid State Electronics, Research, **7**, pp. 209, Jan., 1954; pp. 46-53, Feb., 1954; pp. 94-100, March, 1954.

DEBYE, P. P.,¹ and CONWELL, E. M.⁴

Electrical Properties of N-Type Germanium, Phys. Rev., **93**, pp. 693-706.

DUNN, H. K.¹

Remarks on a Paper entitled "Multiple Helmholtz Resonators", Letter to the Editor, Acous. Soc. Am., J., **26**, p. 103, Jan., 1954.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

⁴ Sylvania Electric Products, Bayside, New York.

ELLIS, W. C.¹ and FAGEANT, J.¹

Orientation Relationships in Cast Germanium, J. Metals, **6**, Pt. 2, pp. 291-294, Feb., 1954.

ELLIS, W. C.¹ and GREINER, E. S.¹

Production of Acceptor Centers in Germanium and Silicon by Plastic Deformation, Letter to the Editor, Phys. Rev., **92**, pp. 1061-1062, Nov. 15, 1953.

FAGEANT, J., see ELLIS, W. C.

FINE, M. E.,¹ VAN DUYN, H.,¹ and KENNEY, NANCY T.¹

Low-Temperature Internal Friction and Elasticity Effects in Vitreous Silica, J. Appl. Phys., **25**, pp. 402-405, March, 1954.

FULLER, C. S., see PEARSON, G. L.

FULLER, C. S., see SEVERIENS, J. C.

GALT, J. K.,¹ YAGER, W. A.,¹ and MERRITT, F. R.¹

Temperature Dependence of Ferromagnetic Resonance Fine Width in a Nickel Iron Ferrite — A New Loss Mechanism, Phys. Rev., **93**, pp. 1119-1120, March, 1954.

GERMER, L. H.¹

Arcing at Electrical Contacts on Closure. Part IV — Activation of Contacts by Organic Vapor, J. Appl. Phys. **25**, pp. 332-335, March, 1954.

GOHN, G. R.,¹ GUERARD, J. P.,¹ and HERBERT, G. J.¹

The Mechanical Properties of Some Nickel Silver Alloy Strips, Proc. A.S.T.M., **54**, Jan., 1954.

GREINER, E. S., see ELLIS, W. C.

GUERARD, J. P., see GOHN, G. R.

¹ Bell Telephone Laboratories.

HAGSTRUM, H. D.¹

Reflection of Ions as Ions or as Metastable Atoms at a Metal Surface, Phys. Rev., **93**, p. 652, Feb., 1954. Abstract of paper presented at Gaseous Electronics Conference Oct. 22-24, 1953.

HEFFNER, H., see CLOGSTON, A. M.

HERBERT, G. J., see GOHN, G. R.

JOHNSON, J. B.,¹ and MCKAY, K. G.¹

Secondary Electron Emission from Germanium, Phys. Rev., **93**, pp. 668-672, Feb. 15, 1954.

KAHN, A. H., see TESSMAN, J. R.

KENNEY, NANCY, see FINE, M. E.

KOCK, W. E.¹

Use of the Sound Spectrograph for Appraising the Relative Quality of Musical Instruments, Letter to the Editor, Acous. Soc. Am., J., **26**, pp. 105-106, Jan., 1954.

KRETZMER, E. R.¹

An Amplitude Stabilized Transistor Oscillator, Proc. I.R.E., **42**, pp. 391-401, Feb., 1954.

LEWIS, H. W.¹

Search for the Hall Effect in a Superconductor — Experiment, Phys. Rev., **92**, pp. 1149-1151, Dec. 1, 1953.

LINVILL, J. G.¹

RC Active Filters, Proc. I.R.E., **42**, pp. 555-564, March, 1954.

MACCOLL, L. A.¹

Geometrical Properties of Two-Dimensional Wave Motion, Am. Math. Monthly, **61**, pp. 96-103, Feb., 1954.

¹ Bell Telephone Laboratories.

MACHLUP, S.¹

Noise in Semiconductors: Spectrum of a Two-Parameter Random Signal, *J. Appl. Phys.*, **25**, March, 1954.

MATTHIAS, B. T.¹

Transition Temperatures of Superconductors, *Phys. Rev.*, **92**, pp. 874-876, Nov. 15, 1953.

McKAY, K. G., see JOHNSON, J. B.

MERRITT, F. R., see ANDERSON, P. W.

MERRITT, F. R., see GALT, J. K.

MORIN, F. J.¹

Lattice-Scattering Mobility in Germanium, *Phys. Rev.*, **93**, pp. 62-63, Jan. 1, 1954.

MORIN, F. J., see PEARSON, G. L.

PEARSON, G. L.,¹ and FULLER, C. S.¹

Silicon p-n Junction Power Rectifiers and Lightning Protectors, *Proc. I.R.E.*, **42**, pp. 760, April, 1954.

PEARSON, G. L.,¹ READ, W. T., JR.,¹ and MORIN, F. J.¹

Dislocations in Plastically Deformed Germanium, *Phys. Rev.*, **93**, pp. 666-667, Feb. 15, 1954.

PFANN, W. J.¹

Comment on Paper by Tiller, Jackson, Rutter and Chalmers, Letter to the Editor, *Acta Metallurgica*, **1**: pp. 763-764, Nov., 1953.

PFANN, W. G.¹

Redistribution of Solutes by Formation and Solidification of a Molten Zone, *J. Metals*, **6**, Pt. 2, pp. 294-297, Feb., 1954.

¹ Bell Telephone Laboratories.

PIERCE, J. R.¹

Coupling of Modes of Propagation, J. Appl. Phys., **25**, pp. 179-183, Feb., 1954.

READ, W. T., JR.¹

Dislocations or What Makes Metals So Weak? Metal Progress, **65**, pp. 101-106, 168, 170, 172, Feb., 1954.

READ, W. T., JR., see PEARSON, G. L.

REMEIKA, J. P., see ANDERSON, P. W.

ROSE, D. J.,¹ and ALLIS, W. P.¹

Transition from Free to Ambipolar Diffusion, Phys. Rev., **93**, pp. 84-93, Jan. 1, 1954.

RYDER, R. M.,¹ and SITTNER, W. R.¹

Transistor Reliability Studies, Proc. I.R.E., **42**, pp. 414-419, Feb., 1954.

SCHLAACK, N. F.¹

Development of the LD Radio System, I.R.E., Trans., P.G.C.S., **2**, pp. 29-38, Jan., 1954.

SEVERIENS, J. C.,¹ and FULLER, C. S.¹

Mobility of Impurity Ions in Germanium and Silicon, Letter to the Editor, Phys. Rev., **92**, pp. 1322-1323, Dec. 1, 1953.

SHIVE, J. N., see SLOCUM, A.

SHOCKLEY, W.,¹ see TESSMAN, J. R.

SITTNER, W. R., see RYDER, R. M.

SLOCUM, A.,¹ and SHIVE, J. N.¹

Shot Dependence of p-n Junction Phototransistor Noise, Letter to the Editor, J. Appl. Phys., **25**, p. 406, March, 1954.

¹ Bell Telephone Laboratories.

SMITH, C. S.¹

Piezoresistance Effect in Germanium and Silicon, *Phys. Rev.*, **94**, pp. 42-49, April 1, 1954.

STILES, K. P.²

Overseas Radiotelephone Services of A. T. & T. Co., *I.R.E., Trans., P.G.C.S.*, **2**, pp. 39-44, Jan., 1954.

STUBBS, R. R.⁵

Telephone Service is a Big Bargain, *Telephony*, **146**, pp. 20-21, 43, March 13, 1954.

TESSMAN, J. R.,⁶ KAHN, A. H.,⁶ and SHOCKLEY, W.¹

Electronic Polarizabilities of Ions in Crystals, *Phys. Rev.*, **92**, pp. 890-895, Nov. 15, 1953.

THOMAS, D. E.¹

Single Transistor FM Transmitter, *Electronics*, **27**, pp. 130-133, Feb., 1954.

VAN DUYNE, H., see FINE, M. E.

VALDES, L. B.¹

Resistivity Measurements on Germanium for Transistors, *Proc. I.R.E.*, **42**, pp. 420-427, Feb., 1954.

WILFONG, J. C., JR.⁷

The Telephone, an Instrument of Culture, *Telephony*, **146**, pp. 22-23, 45, March 20, 1954.

WOJCIECHOWSKI, B. M.³

Capacitance Gage Checks Cable Sheath Thickness, *Electronics*, **27**, pp. 134-137, April, 1954.

YAGER, W. A., see GALT, J. K.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

³ Western Electric Company, Inc.

⁵ Southern Bell Telephone Company.

⁶ Department of Physics, University of California, Berkeley, Calif.

⁷ Chesapeake and Potomac Telephone Company.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

BIELING, C. A., see EDELSON, D.

BOWN, R.

Vitality of a Research Institution and How to Maintain It, Monograph 2207.

CAMPBELL, W. H.

Current Status of Fretting Corrosion, Monograph 2160.

CONWELL, E. M.

High Field Mobility in Germanium with Impurity Scattering Dominant, Monograph 2158.

DACEY, G. C.

Space-Charge Limited Hole Current in Germanium, Monograph 2157.

EDELSON, D., BIELING, C. A., and KOHMAN, G. T.

Electrical Decomposition of Sulfur Hexafluoride, Monograph 2175.

GRAY, MARION C.

Legendre Functions of Fractional Order, Monograph 2164.

GRISDALE, R. O.

The Formation of Black Carbon, Monograph 2161.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

GROTH, W. D., AND SLADE, F. D.

Principles of Tape-to-Card Conversion in the AMA System and Mechanized Billing of AMA Toll Messages, Monograph 2190.

KOHMAN, G. T., see EDELSON, D.

PIERCE, J. R., and WALKER, L. R.

"Brillouin Flow" With Thermal Velocities, Monograph 2191.

PRIM, R. C., see SHOCKLEY, W.

READ, W. T., JR.

Dislocations or What Makes Metals so Weak? Monograph 2211.

ROSE, D. J.

The Transition from Free to Ambipolar Diffusion, Monograph 2205.

SHOCKLEY, W., and PRIM, R. C.

Space-Charge Limited Emission in Semiconductors, Monograph 2156.

SLADE, F. D., see GROTH, W. B.

THOMAS, D. E.

Low-Drain Transistor Audio Oscillator, Monograph 2204.

THURMOND, C. D.

Equilibrium Thermochemistry of Solid and Liquid Alloys of Germanium and of Silicon, Monograph 2210.

WALKER, L. R., see PIERCE, J. R.

WILKINSON, R. I.

Random Picture Spacing with Multiple Camera Installations, Monograph 2188.

Contributors to this Issue

ALBERT L. BLAHA, B.S. in E.E., Polytechnic Institute of Brooklyn, 1950; Bell Telephone Laboratories, 1936-. In 1937 and 1938, Mr. Blaha was in the quartz crystal development shop. Since then he has been primarily concerned with the testing and development of relays. During World War II he worked on magnetostriction type sonar devices.

A. J. BRUNNER, B.S. in M.E., Lewis Institute, 1934; Western Electric Company, 1920-. During Mr. Brunner's early association with Western Electric he was included in an engineering group developing special machines for the manufacture of telephone products. Later he worked on the development of die casting processes. For the past decade his assignments have been in the field of plastic molding. He did notable work in connection with molding 500-type handset parts and is presently engaged in engineering the facilities required to manufacture molded parts for the wire spring relay. He is the holder of numerous patents.

F. HAROLD CHASE, University of Illinois, 1921; Western Electric Company, 1917-1918; Bell Telephone Laboratories, 1921-. Until joining the Power Engineering Group in 1943 he was concerned with the design of carrier system equipment and maintenance practices on toll equipment. Since 1951, he has been developing new uses for transistors in the control of power equipment.

H. E. COSSON, B.S. in M.E., Michigan College of Mining and Technology, 1949; Allis Chalmers Manufacturing Company, 1949-51; A. O. Smith Corp., 1951; Western Electric Company, 1951-. Mr. Cosson served thirty-one months during World War II, fifteen of which were on a naval aircraft carrier. Since joining the development engineering group at Western Electric Company, he has worked on problems associated with straightening wire for the wire spring relay. Junior member A.S.M.E.

THOMAS E. DAVIS, B.S. in E.E., University of Arizona, 1928; Bell Telephone Laboratories, 1928-. He has been concerned with apparatus development projects, including those related to microphones, handsets

and echo suppressors for long telephone lines. During World War II he worked on underwater sound systems for the Navy and was awarded the Naval Ordnance Development Award by the U. S. Navy. Since then he has been working with the wire-spring relay and line concentrator for the No. 5 crossbar system. Member American Institute of Electrical Engineers and Tau Beta Pi.

B. H. HAMILTON, B.S. in E.E., University of Kansas, 1949; Bell Telephone Laboratories 1950-. With the Laboratories he has worked on development of equipment to power the L3 carrier system and has been concerned with fundamental studies of new types of regulated rectifiers. Member American Institute of Electrical Engineers, Tau Beta Pi, Sigma Tau, Sigma Xi, Kappa Eta Kappa.

A. L. QUINLAN, B.S. in E.E., University of Kansas, 1921; Western Electric Company, 1921-. Prior to World War II, Mr. Quinlan worked extensively on the development of manufacturing methods and machines for loading coils. He was granted patents on loading coil case designs and on toroidal coil winding machines and was co-author of an article, *Recent Improvements in Loading Apparatus for Telephone Cables*, published in the A.I.E.E. Journal, Dec. 1947. During the war he had engineering assignments on gun director, precision coil manufacture and vacuum tube projects. Since then he has developed manufacturing facilities and methods for welding precious metal contacts to telephone switching apparatus. Notable among these are the roll welding of contact tape to crossbar switch multiples and the resistance and percussion welding of contacts to wire spring relays. Member A.I.E.E.

WILLIAM SHOCKLEY, B.Sc., California Institute of Technology, 1932; Ph.D., Massachusetts Institute of Technology, 1936; Teaching Fellow, M.I.T., 1932-1936; Bell Telephone Laboratories 1936-1942; Director of Research, Antisubmarine Warfare Operations Research Group, Division of War Research, Columbia University, 1942-1944; Expert Consultant, Office of the Secretary of War, 1944-1945; Bell Telephone Laboratories 1945-. Appointed Director of Transistor Physics Research December 1, 1953, he had directed the group which invented the point-contact transistor. During the past six years he has made many contributions to solid state physics particularly in connection with the transistor. In addition to solid state physics and semiconductors, his work has also included vacuum tube and electron multiplier design, studies of various physical phenomena in alloys, radar development and magnetism. Medal for Merit, U. S.

War Department, 1946; Air Force Association Citation of Honor, 1951; Morris Liebmann Memorial Prize, I.R.E., 1952; Oliver E. Buckley Solid State Physics Prize, American Physical Society, 1953; Certificate of Appreciation, Department of Army, 1953; Comstock Prize, National Academy of Sciences, 1954. Fellow of American Physical Society; Senior Member Institute of Radio Engineers; Member of National Academy of Sciences, Tau Beta Pi, Sigma Xi. For the past few months he has been on leave to California Institute of Technology for teaching and study in the field of solid state physics.

DONALD H. SMITH, B.S. in E.E., University of Minnesota, 1944; Bell Telephone Laboratories, 1947-. After working with the Systems Department of the Laboratories on trial installations, Mr. Smith was concerned with rectifiers and regulating systems in power development. He is currently in charge of the group doing long-range engineering on power development. Member of A.I.E.E. and the Amateur Astronomers Association.

R. W. STRICKLAND, B.M.E., University of Florida, 1951; Western Electric Company, 1951-. Mr. Strickland served two and a half years in the U. S. Armed Forces prior to receiving his degree. Since coming to the Western Electric Company, he has been active in the development of equipment and processes for molding of plastic components of the wire spring relay. Junior member A.S.M.E.

HARRY SUHL, B.Sc., University of Wales, 1943; Ph.D., Oriel College, University of Oxford, 1948. Admiralty Signal Establishment, 1943-46; Bell Telephone Laboratories, 1948-. Dr. Suhl conducted research on the properties of germanium until 1950 when he became concerned with electron dynamics and solid state physics research. His current work is in the applied physics of solids. Member of the American Institute of Physics and Fellow of the American Physical Society.

ERIC E. SUMNER, B.M.E., Cooper Union, 1948; M.A. Degree in Physics, Columbia University, 1953; Instructor of Physics, Cooper Union, 1947-48; Non-resident instructor of Massachusetts Institute of Technology on *Probability and Statistics — Applications to Sampling and Quality Control*, summer, 1950; Bell Telephone Laboratories, 1948-. Mr. Sumner was given rotational assignments in apparatus, switching, and Television transmission development and switching research, and has worked on a number of projects, including the card translator, the mag-

netic drum, video transmission evaluator, vibrating reed selector, development of wire-spring relay, trouble recording apparatus development, and transistor circuitry for a subscriber line concentrator. He is currently engaged in developing small functional circuits for an electronic switching system. Member of Tau Beta Pi and Pi Tau Sigma.

ERLING D. SUNDE, E.E., Technische Hochschule, Darmstadt, Germany, 1926. Brooklyn Edison Company, 1927; American Telephone and Telegraph Company, 1927-1934; Bell Telephone Laboratories, 1934-. Mr. Sunde's work has been centered on theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series Book. Member of the A.I.E.E., the American Mathematical Society, and the American Association for the Advancement of Science.

LAURENCE R. WALKER, B.Sc. and Ph.D., McGill University, 1935 and 1939; University of California, 1939-41. Radiation Laboratory, Massachusetts Institute of Technology, 1941-1945; Bell Telephone Laboratories, 1945-. Dr. Walker has been primarily engaged in research on microwave oscillators and amplifiers. At present he is a member of the physical research group concerned with the applied physics of solids. Fellow of the American Physical Society.

THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII

SEPTEMBER 1954

NUMBER 5

KANSAS CITY, MO.
PUBLIC LIBRARY

OCT 5 1954

Motion of Individual Domain Walls in a Nickel-Iron Ferrite

J. K. GALT 1023

Negative Impedance Telephone Repeaters

J. L. MERRILL, JR., A. F. ROSE, AND J. O. SMETHURST 1055

Stress Systems in the Solderless Wrapped Connection and Their
Permanence

W. P. MASON AND O. L. ANDERSON 1093

A New Multicontact Relay for Telephone Switching Systems

I. S. RAFUSE 1111

Topics in Guided Wave Propagation Through Gyromagnetic Media
Part III—Perturbation Theory and Miscellaneous Results

H. SUHL AND L. R. WALKER 1133

Bell System Technical Papers Not Published in this Journal 1195

Recent Bell System Monographs 1201

Contributors to this Issue 1206

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

S. BRACKEN, *Chairman of the Board,*
Western Electric Company

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. McNEELY, *Vice President, American Telephone*
and Telegraph Company

EDITORIAL COMMITTEE

W. H. DOHERTY, <i>Chairman</i>	F. R. LACK
A. J. BUSCH	W. H. NUNN
G. D. EDWARDS	H. I. ROMNES
J. B. FISK	H. V. SCHMIDT
E. I. GREEN	G. N. THAYER
R. K. HONAMAN	J. R. WILSON

EDITORIAL STAFF

J. D. TEBB, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXIII

SEPTEMBER 1954

NUMBER 5

Copyright, 1954, American Telephone and Telegraph Company

Motion of Individual Domain Walls in a Nickel-Iron Ferrite

By J. K. GALT

(Manuscript received May 11, 1954)

Samples have been cut from single crystals of the nickel-iron ferrite $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$ in such a way that they contain one and only one movable ferromagnetic domain wall. The viscous damping coefficient for this wall, which is a measure of the losses associated with domain wall motion in this material, has been measured as a function of temperature. This damping shows a very large increase as the temperature goes down to the region of 77°K . The value of this damping is correlated with the Landau-Lifshitz equation for the rotational motion of magnetization by means of previously available theoretical analysis. In addition, it is suggested that the sharp increase in damping at low temperatures is due to a relaxation associated with a rearrangement of the valence electrons on the divalent and trivalent iron ions in the ferrite. A tentative phenomenological theory of the losses based on this mechanism is presented.

INTRODUCTION

The mechanism which contributes most to the permeability of high-permeability magnetic materials is the motion of ferromagnetic domain walls. These walls are thin lamellae in which the direction of the spontaneous magnetization of the material changes from one domain to another. As a result, this mechanism contributes a major part of the energy losses which accompany rapid changes in the direction of the

magnetization in such materials. In the ferromagnetic metals, it is well known that these losses ordinarily arise largely from the eddy currents which are induced by the motion of the domain walls. In the ferrites, however, the conductivity is so low that the contribution of eddy currents to the losses is never overwhelming and is often negligible; the losses must therefore in large part arise from other sources not yet understood. It is the purpose of this paper to present some recent studies of these losses and to discuss their relevance to the losses in ferrites generally.

In any ordinary sample of a ferromagnetic material, a study of domain wall motion and the associated energy losses is complicated by the fact that the domain pattern is very complex. Any attempt to provide a theoretical explanation of data taken on such samples must involve an averaging process over many domain walls of varying area, crystal orientation, etc. This makes it extremely difficult to describe the behavior of such patterns uniquely and quantitatively, although some progress has been made.^{1,2} A method of avoiding this difficulty has been developed by Williams, Bozorth, Shockley, Kittel³ and Stewart^{3a} in working on silicon iron. This method consists in cutting a polygonal ring from a single crystal in such a way that each leg of the ring lies along one of the easy directions of magnetization in the crystal. In silicon iron this leads to a rectangular ring with each leg along a [100] crystal direction. In the ferrite which we use this technique to study here, the easy directions are [111] directions, and we use a diamond shaped sample as shown by the solid lines in Fig. 1. Each leg is along a [111] direction, and the major face is a (110) plane. If the sample is good enough, the domain pattern is that indicated by the dotted lines in Fig. 1. This pattern consists of four stationary walls, one at each corner, and one movable wall which goes all the way around the sample. The magnetization thus travels around the sample in two paths, one clockwise and the other counter-clockwise, and the position of the movable wall therefore determines the net circumferential magnetization. In such samples we study quantitatively and in some detail the motion of an individual movable wall.

Williams, Shockley and Kittel³ studied the motion of the movable wall on one of the rectangles cut from a single crystal of silicon iron. They found the motion to be viscously damped, as Sixtus and Tonks⁴ had in earlier experiments with more complicated domain walls. Because of the simplicity of their domain pattern, Williams, Shockley and Kittel were able to calculate the eddy current losses in their experiments, and to show that they accounted for most of the observed damp-

ing as expected. There was an additional contribution, however, which Kittel suggested was due to mechanisms of the sort which give rise to the width of ferromagnetic resonance lines.⁵ These mechanisms of course are the controlling ones in the ferrites, where eddy current losses are small. The motion of a domain wall damped by such effects and unaffected by eddy currents was first discussed in a classic paper by Landau and Lifshitz.⁶

The experiments reported in the present paper consist of measurements of the velocity of a movable wall as a function of applied magnetic field in a sample like that shown in Fig. 1. The measurements are made by observing the voltage induced in a secondary winding on such a sample when a known field is applied by means of a pulse of current in a primary winding. The composition of the ferrites used in these studies is given by the approximate chemical formula $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$. Data have been taken as a function of temperature on several samples. The large, perfect crystals of the ferrites which are essential to the success of these experiments have been obtained through Dr.

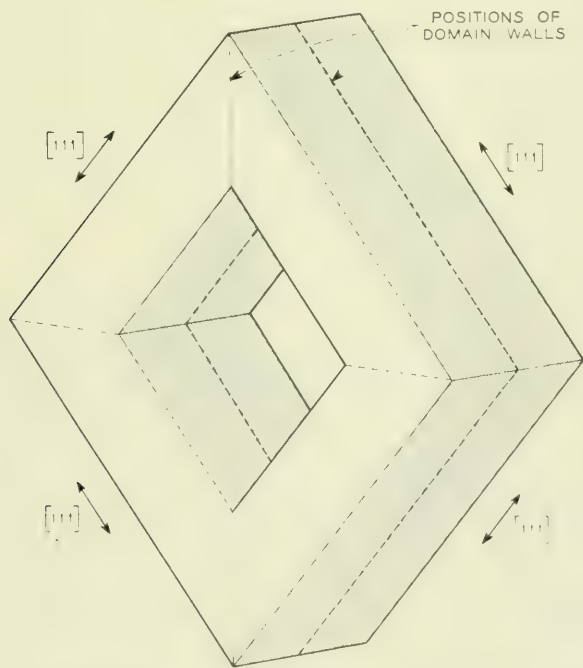


Fig. 1 — Sample of ferrite used in this study.

G. W. Clark from the Linde Air Products Company. Similar studies have been performed previously at room temperature on Fe_3O_4 ⁷ by the author, and in a preliminary way on $(\text{NiO})_{0.91}(\text{FeO})_{0.09}\text{Fe}_2\text{O}_3$ ⁸ by the author in collaboration with J. A. Andrus and H. G. Hopper.

THE EXPERIMENTS

Preparation of Samples

The key to the success of experiments of this sort is of course obtaining the domain pattern shown in Fig. 1, so that we have only one movable wall in the sample. The achievement of this pattern, and the observation of it when achieved, depend in turn on success in producing a perfect or almost perfect sample. It therefore seems worth while to describe the process which has finally emerged as a satisfactory way of producing these samples.

The rough crystal is first oriented by means of X-rays (Laue and X-ray goniometer techniques) to an accuracy of a few minutes while it is mounted in such a position that afterwards we can grind a flat on it which is coincident with the (110) plane. This flat is ground simply with a belt grinder. A cut is then made with a diamond saw parallel to this flat, so that we have a disc whose faces are (110) crystal planes. This disc is usually made about $2\frac{1}{2}$ to 3 mm thick. The second face is ground accurately parallel to the first in a paralleling block. A flat coincident with the (100) plane is ground on the edge of this disc for later use in orienting the sample in the plane of the disc. This flat is also ground with a belt grinder after orienting the disc with Laue and X-ray goniometer techniques.

The disc is ground as smooth as possible with 303¹/₂ emery, and then very carefully polished. The polishing is done first on a lap surfaced with No. 0000 french emery paper, with Linde A abrasive loose on top of the emery paper. After this the lap is surfaced with a sheet of very smooth paper and then Linde B abrasive is used. The polishing process takes four to eight hours per disc, and removes all pits visible under 50 $\times$ magnification, except those inherent in the crystal. Only on such a smooth surface can the very small holes which sometimes occur in these crystals be seen. Sometimes, however, the polishing process conceals fine cracks. It also cannot reveal variations in chemical composition which sometimes occur from point to point in the crystals. Such composition variations are presumably variations in the concentration of divalent iron from point to point in the crystal. In order to reveal these latter imperfections, the disc is etched for three hours by boiling it in

50 per cent H_2SO_4 under a reflux condenser. An asbestos pad is placed between the flame and the bottom of the flask containing the H_2SO_4 in order to prevent sharp temperature fluctuations in the bath. The rate of the etching attack, and the quality of the surface it leaves are sharply dependent upon temperature. It should be mentioned that if this etch is used on the discs *before* polishing, the surface remains rough or may even be made rougher, so that it is impossible to detect the imperfections in the disc. The etch must start on a smooth surface.

Once the disc is cut, polished, and etched, if it is found to be sufficiently free of imperfections, a sample is cut from it in such a way as to avoid those which there are, as they are revealed by the polishing and etching processes. First the diamond-shaped hole is cut by means of a jig whose rotational position with respect to the disc is determined from the (100) flat on the edge of the disc. This jig is a piece of steel which is driven in vibration vertically with a magnetostrictive drive.⁹ The surface of the disc is covered with a slurry of carborundum or diamond dust, and this abrasive is made to cut a hole in the disc as the vibrating jig is slowly lowered. With the hole cut in the proper orientation, the outer parts of the disc are ground down to form the legs of the sample. Another jig of the proper shape is used to hold the sample in position during this process.

A hysteresis loop is taken as soon as the sample is cut. A relatively good loop taken on our best sample is shown in Fig. 2. All such loops on these samples are taken on the Cioffi recording fluxmeter.¹⁰ This loop is obviously not yet in the form which we finally need. In order to square the hysteresis loop, we anneal the sample for approximately an hour at 600°C in a magnetic field of 10 to 20 oersteds. The field is produced by running a current through a few turns of glass insulated wire wound on the sample. After such a heat treatment the hysteresis loop of this sample assumed the form shown in Fig. 3.

Once the sample is prepared, the next problem is to observe the domain pattern and find if any important deviations from the pattern shown in Fig. 1 occur. The heat-treatment we give them corrodes the polished surfaces of the sample, and of course the faces exposed when the sample is cut from the disc have not yet been polished. Consequently both the major (110) faces of the sample and the outer faces of the legs are polished, and the sample is then etched in the same way as before. Usually a hysteresis loop is again taken at this point as a check. If the sample is good, it is not significantly different from the loop taken immediately after heat-treatment. The sample is brought to a demagnetized condition at this point so that the movable wall will be near the center of the sample where it can be observed. This completes the process of

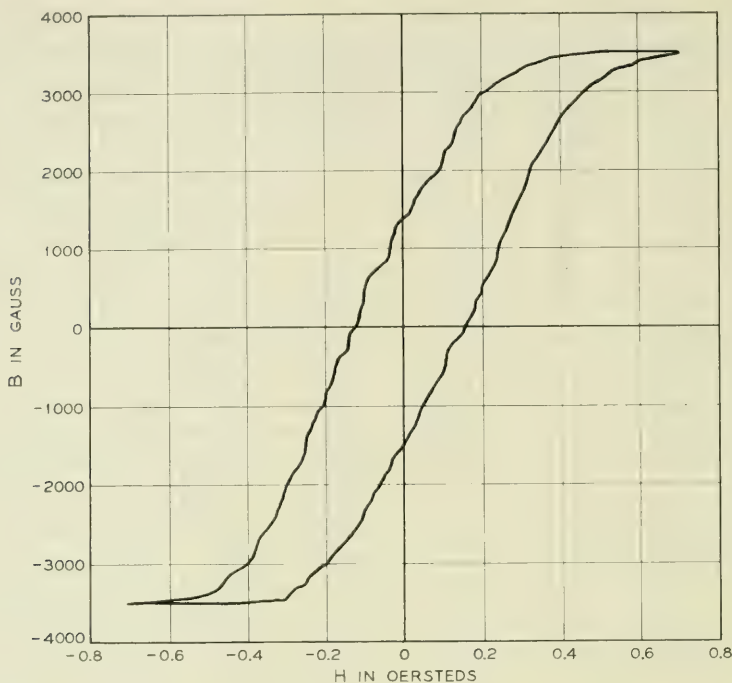


Fig. 2 — Hysteresis loop taken on sample before heat treatment in a magnetic field.

preparing the samples for observing their domain pattern and for performing our experiments on them.

Four samples have been prepared in this way for our studies on $(\text{NiO})_{0.75} (\text{FeO})_{0.25} \text{Fe}_2\text{O}_3$.

Domain Pattern Observations

The method used to observe the domain walls on these surfaces is the same as that used by Williams and his collaborators.³ We will therefore not describe it in detail. It consists essentially of observing through a microscope the pattern formed by a magnetic colloid on the surface.

The observation of domain patterns, even on these carefully prepared samples, is difficult. There is still some pitting on the surfaces. Also, many of the surfaces become rounded in the process of polishing. This produces surface spikes of the sort discussed by Williams, Bozorth and Shockley.³ The result is that on many surfaces the domain pattern of the sample as a whole has to be discerned in a substantial amount

of extraneous structure such as surface spikes and pits. It is therefore impossible to show in one picture the whole pattern as diagrammed in Fig. 1. However, more detailed pictures of parts of the pattern do show that it is there. The essential features of the pattern on our best sample are shown in Figs. 4 and 5. Fig. 4 shows the stationary wall at one corner, and Fig. 5 shows a section of the movable wall on one leg. Differentiation of the domain walls in Figs. 4 and 5 from the many scratches is not very difficult after one has some experience in such observations.

The variation in wall position along the leg shown in Fig. 5 is due to the effects of strains and other imperfections, present even in this carefully prepared sample, in determining the position of the wall at rest.

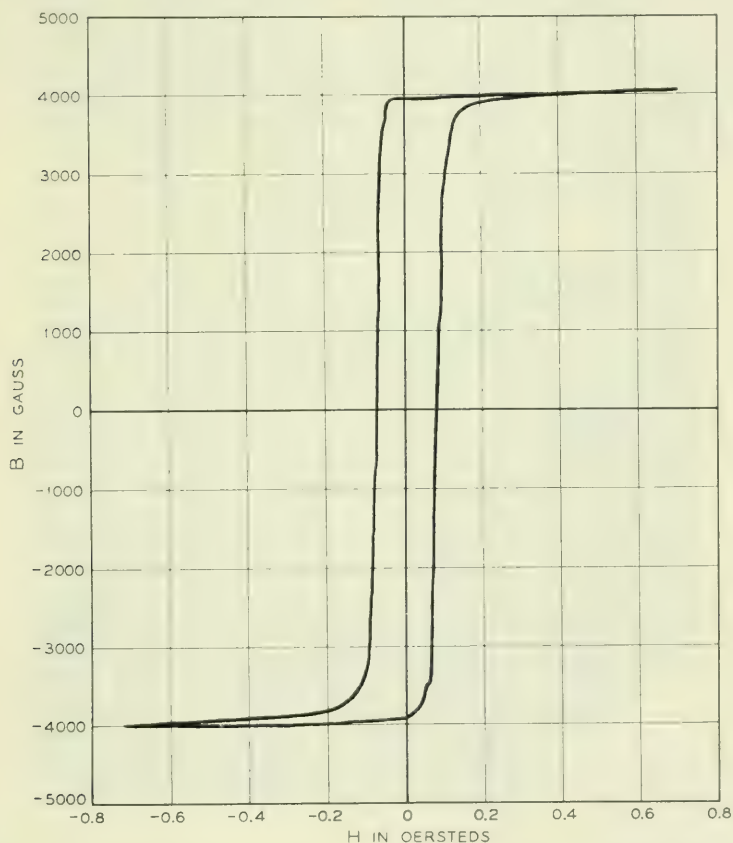


Fig. 3 — Hysteresis loop taken on same sample as loop in Fig. 2 after annealing for one hour at 580°C in a magnetic field of approximately 20 oersteds.

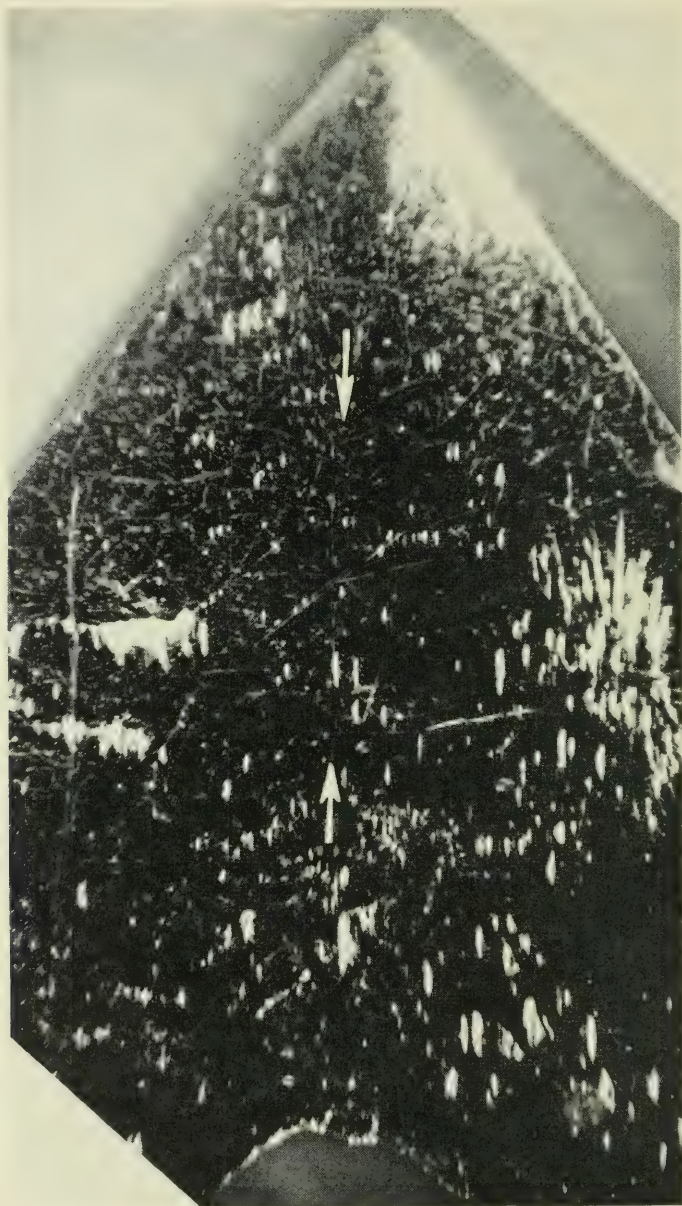


Fig. 4 — Picture of domain pattern at one corner of sample. The stationary wall is indicated by arrows.

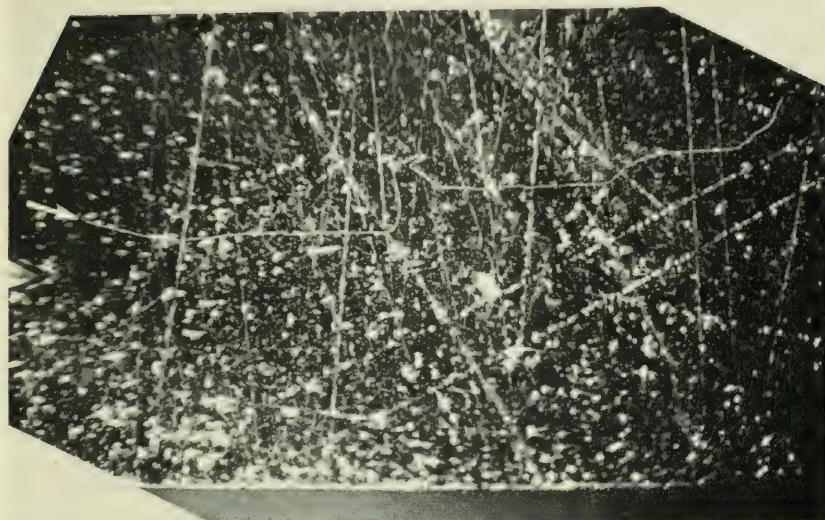


Fig. 5 — Picture of a section of the movable wall along one leg of sample. The line has been slightly emphasized by retouching after the picture was taken. The edge of the leg can be seen at the bottom of the figure.

It is unlikely that the wall is so distorted from a plane when in rapid motion, however, since then the driving force and the viscous damping resistance to motion are both much larger than the effects of these imperfections. The imperfections, of course, are primarily effective in determining the coercive force, as read from the hysteresis loop.

The domain pattern as traced out on each of four samples is discussed below:

Sample 1. Although they had spikes associated with them, the stationary walls expected at the corners could be seen, at least in part. In addition, at one of the acute angle corners there was some rather extensive domain wall structure. This structure had one form when the sample was magnetized in one direction, another when it was magnetized in the other. It was due to the presence of a small void at this corner whose magnetic energy was reduced by having domains of reversed magnetization around it. The existence of this void was established from structure which appeared in the spots on an X-ray Laue photograph taken at this point. The process of magnetization in this sample consisted in (a) the growth of the wall from the nucleus around this void until it existed all around the ring, and (b) the motion of this wall to the other side of the sample, where it again shrank to a configuration

which minimized the magnetic energy of the void. As a result of this state of affairs, the wall was not in equilibrium in the middle of the sample, but always shrank around the void so as to magnetize the sample in one direction or the other. It was impossible therefore to demagnetize the sample so that the wall was at the center of the legs and then have the wall stay there to be observed. The wall could be brought to the center, however, and held there using the technique mentioned by Williams and Shockley³ (see Fig. 8 in their paper). The legs of the sample were so small that it was difficult to do this, but the wall was found on three legs of the sample at different times in this way. The wall curved a good deal in traveling along the legs. This sample was etched repeatedly so that data could be obtained as a function of sample dimensions. The variations observed led to a viscous domain wall damping independent of dimensions if it was assumed that the domain wall was perpendicular to the major (110) face. Therefore it was to this (11 $\bar{2}$) plane that the wall was brought for observation.

Sample 2. Stationary walls at the corners were rather patchy but visible. The movable wall was traced along the outside faces of two legs, which indicates that it lay in the (110) plane. The wall curved a good deal. There were also other walls which enclosed patches of surface. It is suspected that these patches were the bases of spike domains extending into the sample from strain patterns on the surface. The sample was therefore etched again. Unfortunately, the bath apparently became locally overheated, and this etch took off rather more material than expected. It also left a matte surface on which domain walls could not be observed. The data taken on this sample, however, check those on other samples if we assume a pattern in which the movable wall is in the (110) plane, as our observations lead us to suspect.

Sample 3. The stationary walls at the corners were seen, but only with difficulty. They were patchy. There was a good deal of structure all along the legs on the major (110) face of this sample, but no wall which ran around the sample could be seen on this face.

On the outside of two of the legs, which are (11 $\bar{2}$) faces, pitting and extraneous walls were so bad that the main wall could not be discerned. On a third, the wall could be traced most of the way. On the fourth, however, there were two walls, one of which could be traced along the whole leg, the other of which went only three-fourths of the way along the leg. Both walls on this leg showed a good deal of curvature. It therefore appears that the movable wall lies in the (110) plane, but that there is *another wall* big enough so that it may move and affect our data. This picture of a domain pattern with two movable walls was confirmed by checking the data obtained on this sample with those from others.

Sample 4. The stationary walls at the corners, although patchy, were seen. There was some extraneous domain wall structure on the major (110) faces, but nothing which looked at all like the main wall.

On the outside (112) faces of the legs, however, the wall was traced almost all the way around. Only short sections were impossible to trace. The wall curved as usual, and there was some extraneous domain wall structure on these faces, but substantially the whole of the ideal pattern shown in Fig. 1 was seen on this sample. The hysteresis loops shown in Figs. 2 and 3 and the domain pattern pictures shown in Figs. 4 and 5 were taken on this sample.

Not only was the domain pattern on sample 4 the best and most complete, but the data taken on this sample was much the most reproducible. We shall therefore report the data taken on this sample in detail, and simply refer to the results on other samples as a check and to indicate the sort of variations which occurred from sample to sample.

Measurements

Our procedure in making the measurements is as follows. The sample is wound with a primary and a secondary winding. A square pulse of positive voltage is applied to the primary winding in series with a resistor which is large enough to keep the pulse rise time short. The rise time must be short compared to the time required for the field produced by the pulse to reverse the magnetization of the sample. On the other hand, since the pulse is applied for the purpose of reversing the magnetization of the sample, the length of the pulse must be at least comparable with the time required for the reversal to occur; if possible, it should be longer than this. The reversal time, of course, is the time required for the mobile domain wall to move from one side of the sample to the other under the field produced by the applied pulse. A second pulse, of negative voltage, is applied to the primary during each cycle of the pulser in order to bring the wall back to its original position so that the phenomenon may be observed repetitively.

By synchronizing an oscilloscope sweep with the pulser, the signal induced in the secondary winding is observed while the applied pulse is on the primary. Since this signal is proportional to the velocity of the wall, it is constant to a first approximation during the application of a constant field. Irregularities in the crystal may cause the velocity of the wall to vary somewhat as it moves across the sample, however, and this will cause the signal to vary too. In this case, the observer reads the average value. Fig. 6 shows an example of the signal induced in the secondary winding as seen on an oscilloscope. Sample 4 was used to obtain this picture.

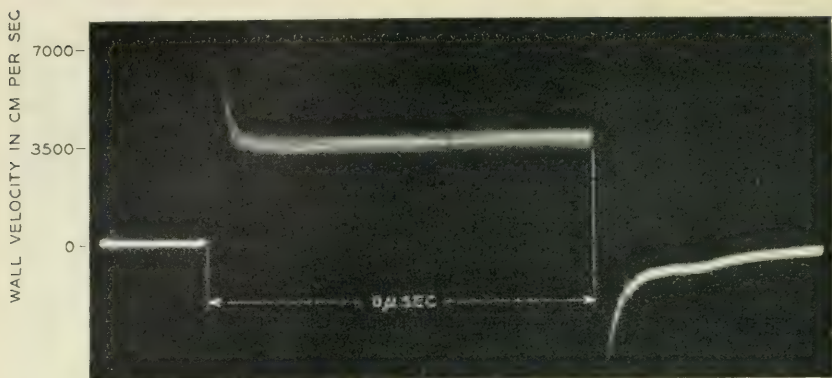


Fig. 6 — Oscilloscope trace of the induced voltage in the secondary winding while a square current pulse 8μ sec long was applied to the primary. The voltage spike at the beginning is not completely understood, but is thought to be connected with the fact that the domain wall spikes associated with imperfections do not pull back on the domain wall until it has moved a little distance. In this sample, once this initial peak is over, the wall velocity comes to its steady state value and is not much disturbed by imperfections, so that the observer need do no averaging. In this case, the wall was moving with a velocity of 3500 cm/sec. At this velocity, the wall was unable to reverse the magnetization in 8μ sec, so the signal ends as the magnetic field goes to zero at the end of the pulse. After the pressure on the wall due to the magnetic field stops, the domain wall spikes associated with imperfections pull the wall back slightly, giving rise to the voltage spike in the opposite direction.

The applied field due to the primary pulse is deduced from the current in the primary winding (measured by observing the voltage across the series resistor) using the solenoid formula $H = 4\pi NI$. To obtain the relation between wall velocity and induced voltage per secondary turn we have:

$$\text{Volts/turn} = (d\Phi/dt) \times 10^{-8} = 8\pi M_s (\Delta z/\Delta t) w_{\text{wall}} \times 10^{-8}, \quad (1)$$

where $(\Delta z/\Delta t)$ is equal to the domain wall velocity v , and w_{wall} is the width of the wall between the boundaries of the sample in the direction perpendicular to the direction of magnetization. It is in deriving (1), of course, that we use our detailed knowledge of the domain pattern in the sample.

We are thus able to obtain a value of the domain wall velocity v for each value of the applied field H . These data are the results of the experiment.

The value of M_s used in (1) is the measured value (322.5 cgs units/cc) at room temperature. The values used at other temperatures have been deduced on the assumption that M_s varies the same way with temperature in this material as it does in magnetite as measured by Weiss and

Forrer.¹¹ We have extrapolated their data to get the variation up to our highest temperatures.

In general, a plot of the data turns out to have the form shown in Fig. 7. This is the data taken on Sample 4 at 201°K. The wall does not move until the field exceeds the coercive force required to get it past various imperfections in the crystal. Its motion in fields higher than this is viscously damped. The wall velocity, v , therefore follows the relation:

$$v = G(H - H_c), \quad (2)$$

where H_c is the coercive field and G is the slope of the line drawn through the data. The value of G is high if the losses are low, and vice versa, of course.

RESULTS

Data on Sample 4 of the sort shown in Fig. 7 have been taken at various temperatures. We show in Fig. 8 a plot of $v/(H - H_c)$ as a function of temperature for this sample. Clearly, the outstanding feature of the data is a tremendous increase in the viscous damping of the domain wall at low temperatures.

Since the other samples were not as satisfactory, for reasons given above, we do not reproduce the data on them explicitly. Similar data have been taken on Samples 1 and 2, however, and they show the same behavior within their accuracy except that the very sharp decrease in $v/(H - H_c)$ seemed to occur at a somewhat higher temperature. This difference may be due to slight variations in composition among the

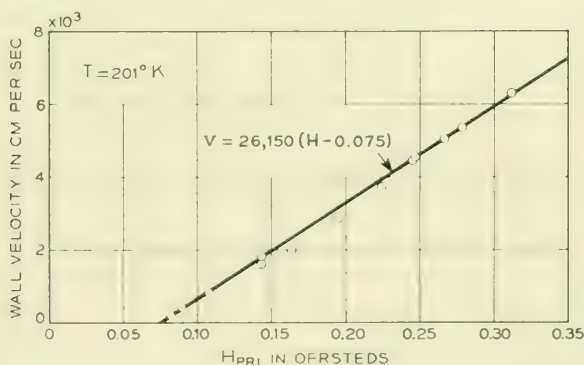


Fig. 7. Typical plot of actual data for domain wall velocity as a function of applied field.

samples (Chemical analysis indicated that Sample 1 was



as distinct from



for Sample 4) but it is possible that other more subtle differences such as the arrangement of the divalent nickel and iron ions are involved. Voltages induced in the secondary winding on Sample 3 for various applied fields were much higher than for the other samples at room temperature. This confirms the presence of more than one wall as indicated by the domain pattern. Therefore no further data were taken on this sample.

Each sample, after it had been cooled to the temperature of liquid

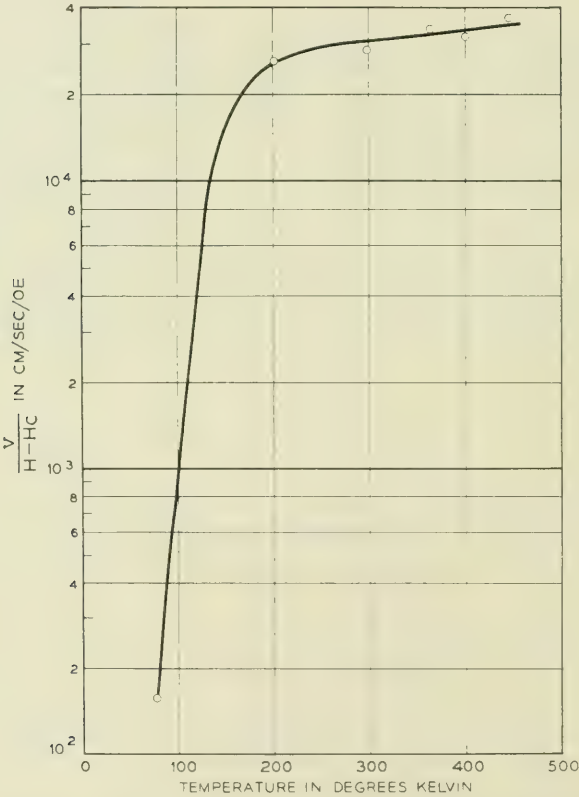


Fig. 8 — Plot of $v_p/(H - H_c)$ for Sample 4 as a function of temperature.

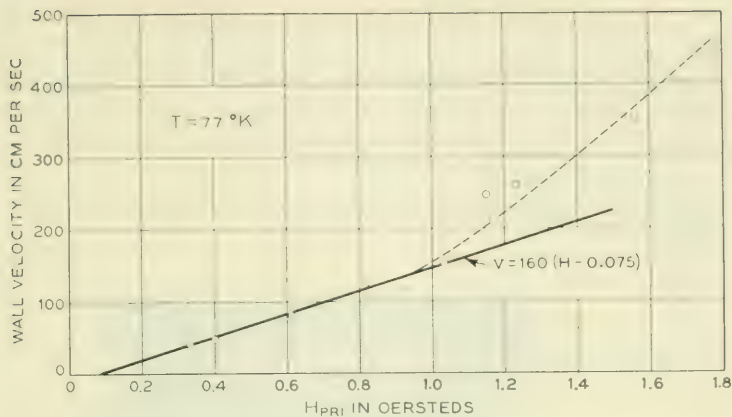


Fig. 9 — Plot of wall velocity as a function of applied field at 77°K. Note that at the higher fields the velocity is no longer a linear function of the applied field.

air once, gave the same value of $v/(H - H_c)$ as before. Samples 1 and 2, however, were cooled several times, and after the later runs they no longer did this. In both cases the value of $v/(H - H_c)$ as deduced from the ideal domain pattern and (1) was lower by about one-half; we interpret this to mean that the domain pattern in these samples was changed by the repeated thermal shock. Direct confirmation of this interpretation by observation was not possible, however, for reasons indicated above in connection with domain pattern observations on these samples.

In view of the fact that only one point is plotted beyond the knee of the curve in Fig. 8, it should be emphasized that continuous qualitative observations made while the sample was cooling showed that the change was continuous and monotonic. On Sample 1, furthermore, the data at 201°K was somewhat down from the knee of the curve [$v/(H - H_c) = 18000$ cm/sec/oe] because of the fact that the knee occurred at higher temperature as mentioned above.

Another feature of the data is indicated by Figs. 9 and 10. Data discussed thus far have been taken at the lowest convenient velocities in order to minimize the possibility of wall distortion. However, when data were taken at higher fields, a non-linearity of the sort shown in Fig. 9 appeared. The average velocity increases more rapidly at the higher fields than our viscous damping coefficient would lead us to expect. This effect was observed at all temperatures, but as comparison of Figs. 7 and 9 will show, it set in at lower velocities at low temperatures where the enlarged viscous damping appeared. Simultaneously with the appearance of this non-linearity in the apparent average velocity of the

wall, the voltage induced in the secondary winding during the pulse of constant applied field is seen from the oscilloscope trace to distort with time. Fig. 10 shows a series of traces observed at room temperature on Sample 4 which show this distortion increasing from (a) to (d) as the applied field is increased. At the highest fields the trace forms a peak which is almost triangular in shape.

We shall discuss the theoretical implications of these data in the next two sections.

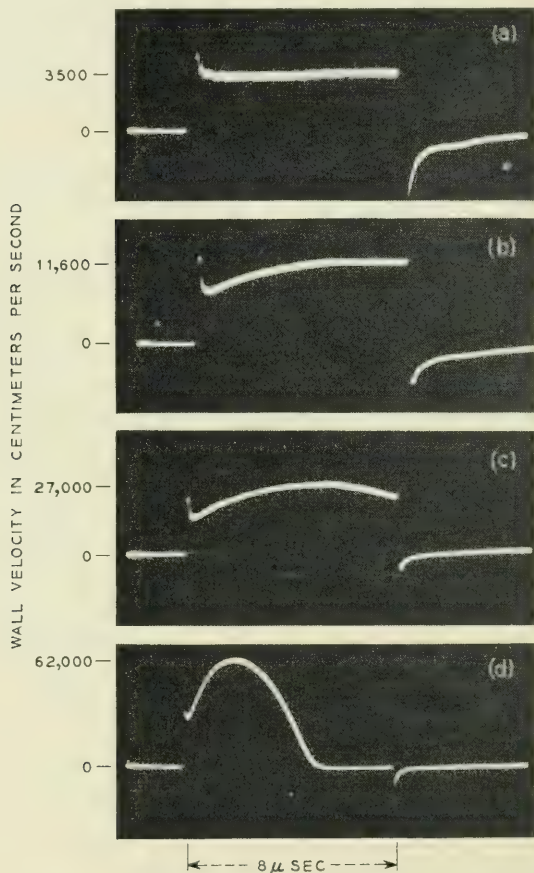


Fig. 10 — A series of oscilloscope traces showing the deviation of the wall velocity from a constant with time at the higher applied magnetic fields and therefore at the higher velocities. These pictures were taken at room temperature, but similar phenomena occur at all other temperatures in the range of applied fields where the nonlinearity in velocity shown in Fig. 9 becomes apparent. The velocities, as shown, increase from *a* to *d*.

THEORY

A theoretical analysis of the experimental results given in the last section divides itself rather naturally into three parts. First we characterize the data in terms of an equation of motion for unit area of domain wall. This means we determine the constants of motion (viscous resistance and coercive force in our case) of unit area of wall. Secondly, we show the relation between the viscous resistance of unit area of domain wall, and the constants which characterize the ferromagnetic material in general (saturation magnetization, crystal anisotropy, etc.). This is essentially an application of the recent work of Becker¹² and Kittel.¹³ Lastly, we calculate the magnitude of the damping from a relaxation mechanism which accounts for the low temperature effect shown in Fig. 8 at least qualitatively.

Consider unit area of a 180° domain wall between two regions of saturated material. Such a system has an equation of motion for small amplitudes of the applied magnetic field H which may be written:

$$m\ddot{z} + \beta\dot{z} + \alpha z = 2M_s H, \quad (3)$$

where z is the displacement of the domain wall along its normal, m is its mass per unit area, β is a parameter measuring viscous resistance, and α is a stiffness parameter which has meaning only for small fields such as those used in initial permeability measurements. When fields larger than the coercive force are applied, as in our experiments, the term containing α disappears and the field effective in moving the wall is less than the applied field by an amount equal to the coercive force; this is shown by the data given previously in the section on results. This reformulation of (3) is quite reasonable when one remembers the spikes which pull back on the wall, in the experiments of Williams and Shockley, for small wall motions and snap off entirely if the wall moves a large distance. Furthermore, since the velocity of the domain wall rises to its steady value in a negligible time in these experiments (Fig. 6) the initial term in (3) is also negligible. As these remarks indicate, under the conditions of the experiment in wall velocity, (2) takes the form:

$$\beta\dot{z} = 2M_s(H_{\text{app}} - H_c). \quad (4)$$

This relation obviously fits the data given previously.

The second step in the theoretical analysis starts from an equation of motion for the magnetization $\vec{M}$ in a small volume which was first used by Landau and Lifshitz,⁶ and takes advantage of more recent work of Becker¹² and Kittel.¹³ If we consider a volume small enough so that the

magnetization in it is everywhere uniform even if we are inside the domain wall, our equation of motion is:

$$\frac{d\vec{M}}{dt} = \gamma[\vec{M} \times \vec{H}] - \lambda/M^2[\vec{M} \times (\vec{M} \times \vec{H})]. \quad (5)$$

Here γ is the gyromagnetic ratio ($gc/2\text{ me}$), and λ is a parameter, assumed to be characteristic of a given ferromagnetic material, which is determined by the magnitude of the damping effects in the motion of $\vec{M}$. The magnitude of the last term on the right in (5) is thus determined by the amount of the damping losses.

The rate of dissipation of energy in the small volume is $\vec{H} \cdot (d\vec{M}/dt)$, where $\vec{M}$ is the magnetic moment of the volume, and $\vec{H}$ is the total magnetic field in the volume. The value of $\vec{H}$ requires some discussion. Outside the wall $\vec{H} = \vec{H}_0$ where $\vec{H}_0$ is the applied field, which is parallel to the wall. When the wall is moving, however, there is an additional field $\vec{H}_e$ inside it. This field, which is normal to the wall, is a demagnetizing field which arises from the tendency of $\vec{M}$ in the moving wall to have a component normal to the wall. This field has just such a value that the magnetization in the moving wall precesses about it with the Larmor frequency. Its value is:

$$\vec{H}_e = -(\vec{v}/\gamma)(\partial\theta/\partial z), \quad (6)$$

as Becker¹² has shown. Here $\vec{v}$ is the velocity of the wall, z is a distance coordinate normal to the wall, and θ is the rotational angle of the magnetization as we pass through the wall along z . Inside the wall, $\vec{H}_e$ is much larger than $\vec{H}_0$, but in any case $\vec{H} = \vec{H}_e + \vec{H}_0$. From (5) we find:

$$\vec{H} \cdot d\vec{M}/dt \cong \lambda H_e^2, \quad (7)$$

as Kittel¹³ first showed. In the theory of the domain wall it is shown that $\partial\theta/\partial z$ in the wall is equal to the square root of the ratio of the increase in anisotropy energy as the magnetization turns away from the easy direction of magnetization, to the exchange energy constant. That is¹⁴:

$$\frac{\partial\theta}{\partial z} = ([g(\theta) - g(\theta_0)]/A)^{1/2}. \quad (8)$$

The exchange energy constant A is defined by the following expression for the exchange energy per unit volume due to gradients in the direction

of magnetization:

$$\text{Exchange energy/unit vol} = A[(\nabla\alpha_1)^2 + (\nabla\alpha_2)^2 + (\nabla\alpha_3)^2], \quad (9)$$

where α_1 , α_2 , and α_3 are the direction cosines of the magnetization. $g(\theta)$ is the anisotropy energy:

$$g(\theta) = K_1 (\alpha_1^2 \alpha_2^2 + \alpha_2^2 \alpha_3^2 + \alpha_3^2 \alpha_1^2), \quad (10)$$

expressed in terms of θ , and $g(\theta_0)$ is the anisotropy energy along the direction of easy magnetization. Note that $[g(\theta) - g(\theta_0)]$ is always positive. K_1 is the first order anisotropy constant.

If we use (6) and (8) in (7), and integrate over z along a cylinder of unit cross-section normal to the wall to get the rate of energy dissipation for unit area of moving wall, we have:

$$\int_{-\infty}^{\infty} \vec{H} \cdot \frac{d\vec{M}}{dt} dz = (\lambda v^2 / \gamma^2 A^{1/2}) \int_{\theta_1}^{\theta_2} [g(\theta) - g(\theta_0)]^{1/2} d\theta = 2H_0 M_s v, \quad (11)$$

where θ_1 and θ_2 are the angular positions of $\vec{M}$ on the two sides of the wall. $\theta_2 - \theta_1 = \pi$, of course, since we are considering a 180° domain wall. In order to obtain (11), we have used (8) to transform from integration over z to integration over θ as well as to evaluate (7). We set our result equal to $2M_s H$ (pressure on the wall) times v since this is the rate at which the wall, considered phenomenologically, does work. We may now write:

$$v = \frac{2M_s \gamma^2 A^{1/2}}{\lambda \int_{\theta_1}^{\theta_2} [g(\theta) - g(\theta_0)]^{1/2} d\theta} H_0. \quad (12)$$

This is the desired relation between wall velocity and applied field which is to be compared with (4). In this way we find:

$$\beta = (\lambda / \gamma^2 A^{1/2}) \int_{\theta_1}^{\theta_2} [g(\theta) - g(\theta_0)]^{1/2} d\theta. \quad (13)$$

We have thus shown the relation between the wall parameter β and the parameter λ which measures in general the losses associated with motions of the magnetization.

The third part of our theoretical analysis is concerned with a calculation of the damping parameter λ , or rather the relation between v and H itself, from an explicit physical mechanism. Such a calculation has not been made in the past since the appropriate mechanism on which to base it has remained obscure. Verwey and his co-workers¹⁵ have explained the well known transition at about 115°K in Fe_3O_4 as

an order-disorder transition in the arrangement of the divalent and trivalent iron ions. M. Fine of Bell Telephone Laboratories has found a remnant of this transition in crystals of the same composition as those studied in the present research, by means of ultrasonic measurements of elastic constants.¹⁶ We propose that the mechanism which causes the sharp rise in the damping of domain wall motion at low temperatures is a relaxation associated with this transition. Wijn and van der Heide¹⁷ have explained in this way observations of their losses associated with initial permeability at low frequencies in certain other polycrystalline ferrites. The time associated with this relaxation should be short because this rearrangement of ions involves only the motion of electrons from one site to another.¹⁸ It should be of the order of the relaxation time associated with the electrical conductivity of Fe_3O_4 . Snoek¹⁹ has suggested some time ago that losses in the ferrites were due to an after-effect (relaxation) which, because of the short time constant involved, must be associated with electron migrations.

It is extremely useful to compare our data with a theory of the damping based on the above relaxation mechanism no matter what assumptions we make in detail about what it is that relaxes. We shall see that we are led quite generally to the result that $v/H \sim 1/\tau$ where τ is the relaxation time for the process. However, in order to perform this calculation explicitly we must make more detailed assumptions about exactly what quantity relaxes with the relaxation time τ . Changes in the direction of the magnetization cause changes in stress in the sample because of magnetostriction. One possible assumption is that the resulting strain would lag behind this stress and mechanical energy would be dissipated in the crystal. This mechanism, however, cannot act in our case. The magnetization in the two domains on each side of the wall points in opposite directions, but causes the same strain in both, and they have such a large stiffness that the thin region occupied by the domain wall assumes this same strain even though the direction of the magnetization is different there. Thus regardless of the stresses produced in the wall, the strains remain the same, inside and outside the wall, whether it is moving or not; under these conditions no work is done on the lattice, and no energy can be lost in this way by the moving wall. A calculation has been made by the author on the assumption that it is the magnetization itself which relaxes with relaxation time τ . This assumption leads to the result that wall velocity is *not* linearly dependent upon $(H - H_c)$; it is therefore not correct, since the data shows such a linear dependence. A similar result is to be expected if we assume that the dielectric polarization relaxes.

What seems at present likely to be the approximate nature of the mechanism, and what we will assume is the nature of the mechanism for the purposes of an illustrative calculation is as follows. As the domain wall passes a point in space, and the direction of $\vec{M}$ changes, the electrons on the divalent and trivalent iron ions tend to rearrange themselves so as to minimize the magnetocrystalline anisotropy energy.²⁰ If $\vec{M}$ changes slowly this anisotropy energy is near the minimum possible value (the *reversible* value) at all times, and the process is almost isothermal. As a result of the fact that the process deviates irreversibly from equilibrium, however, net work is done in bringing about the change. If, on the other hand, the direction of $\vec{M}$ changes so suddenly that the electrons have no time to rearrange, the process is adiabatic, and the magnetocrystalline anisotropy energy varies more widely with the angular position of $\vec{M}$.

Since our data is taken at low velocities and extrapolated to zero velocity, it seems most appropriate for us to make a calculation of the losses on the assumption that as we increase the velocity of the wall we are deviating from the isothermal condition. Let us define as a thermodynamical system the part of the magnetic lattice which lies in a small volume fixed in space. This volume is a sheet of unit cross-section in which the magnetization is uniform and which is part of the cylinder of unit cross-section normal to the wall mentioned in connection with (11). From the first law of thermodynamics, as the wall passes the small volume, we have:

$$dw = dU - dQ = dg, \quad (14)$$

where dQ is heat added to the system, dU is a change in internal energy, dw is work done on the system, and g is the anisotropy energy associated with our rearranging electrons. Note that g , is a term in the free energy of the magnetic lattice. The free energy includes other terms in addition to g , and indeed there is in general another term in the magnetocrystalline anisotropy energy; we will not consider any of these, however, since they also integrate to zero as the domain wall passes our small volume. Similarly, there are many contributions to dw , but we will concern ourselves only with the term which does net work on our system, the pressure on the wall times its velocity. If we now consider changes in our system with time we have from (14):

$$\frac{dw}{dt} = \frac{dg}{dt}. \quad (15)$$

The rate of doing work on unit area of the wall moving along our

cylinder, which is the integral of dw/dt over the cylinder, is $2M_s H_0 v$ as in (11). The rate of dissipation of energy in this section of the wall is somewhat more difficult to calculate. It is the sum of contributions from all the small volumes along the cylinder. In order to calculate the contribution from the small volume we are considering, we first note that if the process is reversible,

$$dg = \frac{dg}{d\theta} d\theta, \quad (16)$$

where θ is the angle of rotation of $\vec{M}$. Physically, the factor $dg/d\theta$ represents a torque on the magnetization. We will assume that (16) holds even when the domain wall passes our system at a finite velocity and we have departed slightly from isothermal equilibrium. The field H_e defined in (6) transmits this torque to make the magnetization rotate through the angle θ , but we will not go further into the details of this process.

As the domain wall goes by, $dg/d\theta$, which we will abbreviate as g' , changes. If the process were reversible, g' would be zero at all times and the torque on the magnetization would always have its equilibrium value. Actually, as we deviate more and more from the reversible process by moving the wall faster, the electrons are no longer able to rearrange fast enough, and g' deviates from zero while continually relaxing toward it. We must form our analysis in such a way that a maximum is established for g' , since even if the magnetization moves infinitely rapidly, g' does not become infinite. Since the electrons minimize their free energy, it must be positive, and we write $g = g_{1\infty} - g_1$ so that $g' = g'_{1\infty} - g'_1$, where g'_1 relaxes toward $g'_{1\infty}$. Note that the torque relaxes *downward* from a value, $g'_{1\infty}$, associated with the adiabatic anisotropy energy (fast motions of the magnetization) to a value zero, associated with the isothermal anisotropy energy (slow motions of the magnetization). We may now write:

$$\frac{dg'_1}{dt} = \frac{g'_{1\infty} - g'_1}{\tau}. \quad (17)$$

Physically, this relation assumes that the torque on the magnetization relaxes in the same way if $g'_{1\infty}$ is a continuous function of time as it would if $g'_{1\infty}$ and g'_1 differed but $g'_{1\infty}$ was constant. Here τ is the relaxation time associated with the rearrangement of divalent and trivalent iron ions. If we write (17) in the form:

$$\frac{dg'_1}{dt} + \frac{g'_1}{\tau} = \frac{g'_{1\infty}}{\tau}, \quad (18)$$

we see at once that the solution for g'_1 as a function of t as the domain wall passes the position of our small volume is:

$$g'_1(t) = \frac{1}{\tau} e^{-t/\tau} \int g'_{1\infty}(t) e^{t/\tau} dt. \quad (19)$$

We have ignored the solution of the homogeneous equation, as it contains a factor $e^{-t/\tau}$.

In order to evaluate (15) for our small volume, we need a more explicit value for g'_1 . From the Fourier transformation we may write:

$$g'_{1\infty}(t) = \int_{-\infty}^{\infty} G'_{1\infty}(\omega) e^{i\omega t} d\omega. \quad (20)$$

Consequently:

$$g'_1(t) = \frac{1}{\tau} e^{-t/\tau} \iint_{-\infty}^{\infty} G'_{1\infty}(\omega) e^{(1/\tau + i\omega)t} d\omega dt. \quad (21)$$

This can be integrated over t . If we do this, then bring the factor $e^{t/\tau}$ out from under the integral sign, we find:

$$g'_1(t) = \int_{-\infty}^{\infty} \frac{G'_{1\infty}(\omega)}{1 + i\omega\tau} e^{i\omega t} d\omega. \quad (22)$$

Now the frequencies for which we get a contribution to the integral in (22) have an upper bound determined by the velocity and the thickness of the domain wall, of course. The faster the domain wall moves, the higher the upper bound on the frequency. We have been careful to make our measurements near zero wall velocity, and the data clearly extrapolates linearly to zero. This was done in order to minimize the possibility of wall distortion, but it also is consistent with our assumption of almost isothermal conditions. We are therefore justified in assuming that $\omega\tau \ll 1$ for all the frequencies in $G'_{1\infty}(\omega)$, and we can expand the factor $1/(1 + i\omega\tau)$ in (22). We obtain:

$$g'_1(t) = \int_{-\infty}^{\infty} [1 - (i\omega\tau) + (i\omega\tau)^2 + \dots] G'_{1\infty}(\omega) e^{i\omega t} d\omega, \quad (23)$$

and if we form the derivatives of (20) and compare them with the terms in (23), we see that:

$$g'_1(t) = g'_{1\infty}(t) - \tau \frac{dg'_{1\infty}(t)}{dt} + \tau^2 \frac{d^2 g'_{1\infty}(t)}{dt^2} + \dots \quad (24)$$

Since the value of τ is very much less than one, the series in (24) converges quite rapidly at low domain wall velocities, and we only keep the first two terms.

Until now, we have been concerned with a small volume of the magnetic lattice fixed in space. As in deriving (11), if we wish to calculate the rate of loss of energy for unit area of the domain wall moving with constant velocity, we must integrate dg_1/dt over a cylinder of unit cross-section normal to the wall. $g'_{1\infty}$ is now a function of $(t - z/v)$ where v is the velocity of the domain wall, and z is a coordinate normal to the wall. If we form this integral, and use (16) and (24), we find:

$$\int_{-\infty}^{\infty} \frac{dg_1}{dt} dz = \int_{-\infty}^{\infty} \left[g'_{1\infty}(t - z/v) - \tau \frac{dg'_{1\infty}}{dt}(t - z/v) \right] \frac{d\theta(t - z/v)}{dt} dz \quad (25)$$

We note that:

$$\begin{aligned} \frac{dg'_{1\infty}(t - z/v)}{dt} &= -v \frac{dg'_{1\infty}(t - z/v)}{dz}, \\ \frac{d\theta(t - z/v)}{dt} &= -v \frac{d\theta(t - z/v)}{dz}. \end{aligned} \quad (26)$$

For the first term in (25) we find, using (26):

$$\int_{-\infty}^{\infty} g'_{1\infty}(t - z/v) \frac{d\theta(t - z/v)}{dt} dz = -v \int_{\theta_1}^{\theta_2} g'_{1\infty}(t - z/v) d\theta = 0, \quad (27)$$

since $g_{1\infty}$ is the same on both sides of the domain wall. Now if we use (26) in the second term of (25), and remember the result in (27), we have:

$$\int_{-\infty}^{\infty} \frac{dg_1}{dt} dz = -\tau v^2 \int_{-\infty}^{\infty} \frac{dg'_{1\infty}(t - z/v)}{dz} \frac{d\theta(t - z/v)}{dz} dz. \quad (28)$$

We may, without loss of generality, evaluate the intergral in (28) at $t = 0$. It is therefore the integral of a function of z over z . In order to evaluate it, we wish to express $g'_{1\infty}$ as a function of θ . We have:

$$\int_{-\infty}^{\infty} \frac{dg_1}{dt} dz = -\tau v^2 \int_{-\infty}^{\infty} \frac{dg'_{1\infty}}{d\theta} \left(\frac{d\theta}{dz} \right)^2 dz, \quad (29)$$

$$= -\tau v^2 \int_{\theta_1}^{\theta_2} \frac{d^2 g'_{1\infty}}{d\theta^2} \frac{d\theta}{dz} d\theta, \quad (30)$$

where we have replaced $g'_{1\infty}$ by $dg'_{1\infty}/d\theta$. We note now, from the relation between g and g_1 , that $dg_1/dt = -dg/dt$. We therefore set the right hand

side of (30) with reversed sign equal to:

$$\int_{-\infty}^{\infty} \frac{dw}{dt} dz = 2M_s H_0 v,$$

as mentioned after (15) to get the relation between the velocity of the domain wall and the applied field. As we shall see later, the right hand side of (30) is positive if we use a $g_{1\infty}$ associated with a positive contribution to the anisotropy energy. We find:

$$v = \frac{1}{\tau} \int_{\theta_1}^{\theta_2} \frac{d^2 g_{1\infty}}{d\theta^2} \frac{d\theta}{dz} d\theta H_0. \quad (31)$$

The derivatives of θ with respect to z in (29), (30), and (31) are to be evaluated by means of (8) and (10). In using these equations, of course, we are assuming that the wall is moving slowly enough so that its shape remains that of the wall at rest.

Equation (31) shows that $v \sim H/\tau$ as we mentioned earlier. Inspection of Fig. 8 shows that this relationship explains very satisfactorily the sharp drop in v ($H = H_c$) at low temperatures if we remember that τ depends on temperature as follows:

$$\tau = \tau_\infty e^{\epsilon/RT}, \quad (32)$$

where ϵ is an activation energy.

Finally, the right hand side of (30) may be set equal to:

$$\int_{-\infty}^{\infty} (\dot{H} \cdot d\vec{M}) dt dz,$$

as calculated from the Landau-Lifshitz equation, see (11), to obtain a value for λ . We find:

$$\lambda = \tau \frac{\gamma^2 A^{1/2} \int_{\theta_1}^{\theta_2} \frac{d^2 g_{1\infty}}{d\theta^2} \frac{d\theta}{dz} d\theta}{\int_{\theta_1}^{\theta_2} [g(\theta) - g(\theta_0)]^{1/2} d\theta}. \quad (33)$$

DISCUSSION

It is clear that (4) fits our experimental data at each temperature. By fitting our data to this expression we obtain values for β , the parameter characteristic of the material which measures the damping of the wall. Values of β obtained in this way are given in Table I. A more

correct value of M_s for Fe_3O_4 at room temperature (475 c.g.s. units) has been used in calculating the β given in Table I than has been used in previous calculations.

In Table I we give in addition to values of β , values of λ at room temperature obtained from (13). The value for Fe_3O_4 is calculated from the new value of β and also corrected for the error mentioned in Reference 14. Values given in References 7 and 8 are somewhat in error. The values of β and λ given for Fe_3O_4 in Table I are the remainders after the contribution due to eddy currents has been subtracted out by means of the low field calculation of Williams, Shockley, and Kittel,³ [see their Equation (11)].

TABLE I—ROOM TEMPERATURE DATA ON Fe_3O_4 AND $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$

	(c.g.s. units)		
	β (corrected for eddy currents)	λ domain wall	λ ferromagnetic resonance
Fe_3O_4	0.44	5.5×10^8	9×10^8
$(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$	0.023	4.6×10^7	10×10^7

TABLE II—DATA FOR $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3^*$

T (°K)	M_s (c.g.s. units)	K_1 (ergs/cc)	$\tau/(H - H_c)$ (cm/sec/oer)	λ domain wall (c.g.s. units)
77	341	-8.1×10^4	158	6.3×10^9
201	335	-5.4×10^4	26150	4.4×10^7
300	322.5	-3.8×10^4	28500	4.6×10^7
363	309	-3.1×10^4	32500-35300	$4.0-4.3 \times 10^7$
400	298	-2.8×10^4	31750	4.5×10^7
445	281	-2.4×10^4	36850	3.9×10^7

* The value of M_s at 300°K is measured from the hysteresis loop of Fig. 3. Other values were obtained by assuming that M_s varied with T in the same way as observed by Weiss and Forrer¹¹ in Fe_3O_4 .

The evaluation of λ from (13) is done as follows. Since the wall is in a (110) plane, we find from (10):

$$g(\theta) - g(\theta_0) = \frac{|K_1|}{4} (2/\sqrt{3} - \sqrt{3} \sin^2 \theta)^2, \quad (34)$$

where θ is the angle between $\vec{M}$ and the [100] direction which lies in the plane of the domain wall. θ is $\cos^{-1}(1/\sqrt{3})$ on one side of the wall, and

$\pi + \cos^{-1} 1(\sqrt{3})$ on the other. Then,

$$\int_{\cos^{-1}(1/\sqrt{3})}^{\pi + \cos^{-1}(1/\sqrt{3})} [g(\theta) - g(\theta_0)]^2 d\theta = 0.92 \sqrt{|K_1|}. \quad (35)$$

In performing this integration, care must be taken to use the positive value of the square root over the whole interval. It should be noted that in using (8) and (10) to evaluate (13) we are assuming that the wall is moving slowly enough so that its shape is the same as that of the wall at rest. A is best evaluated from a fundamental relation derived by Herring and Kittel²¹ between A and the Bloch constant:

$$A = [S_0/\Omega]^{1/3} [k/13.3C^{2/3}], \quad (36)$$

where k is Boltzmann's constant, C is Bloch's constant as used in the relation $M_s = M_0(1 - CT^{3/2})$, S_0 is the atomic spin, and Ω is the atomic volume. (S_0/Ω) is equal to the saturation magnetization at 0°K divided by twice the Bohr magneton.

For Fe_3O_4 , we find $C = 4 \times 10^{-6}$ by fitting the Bloch $T^{3/2}$ law to the saturation magnetization measurements of Weiss and Forrer.¹¹ From (36), assuming M_s at 0°K is 505 c.g.s. units, we then find $A = 1.24 \times 10^{-6}$. Furthermore, $K_1 = -1.1 \times 10^5$ as given by Bickford,²² and

$$\gamma = (1.76 \times 10^7)g/2 = 1.865 \times 10^7,$$

where we have used Bickford's²² value (2.12) of g .²³ Now from (13) we find

$$\lambda = 5.5 \times 10^8$$

in Fe_3O_4 at room temperature.

For $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$, we assume that M_s , while different from that for Fe_3O_4 , varies in the same way with temperature so that $C = 4 \times 10^{-6}$. From our measurement of M_s at room temperature (322.5 c.g.s. units) and this assumption about the variation of M_s with T , we find that M_s at 0°K is 342 c.g.s. units. This leads by (36) to $A = 1.09 \times 10^{-6}$. Ferromagnetic resonance experiments²⁴ done by W. A. Yager and F. R. Merritt in collaboration with the author on spherical single crystals of the same ferrite material as that used in the present research give a g value of 2.14 at all the temperatures mentioned in Table II except 77°K, where $g = 2.19$. These values are used in determining the value of $\gamma = 1.76 \times 10^7 (g/2)$ in (13). The anisotropy values in Table II are also taken from the results of these ferromagnetic resonance experi-

ments. The K_1 at room temperature is -3.8×10^4 . The room temperature value of λ domain wall given in Table I and the values at various temperatures given in Table II are obtained from (13) using these data.

An independent value of λ may be obtained from the ferromagnetic resonance line width. The relation between the observed line width $2\Delta H$ and λ has been given elsewhere.²⁵ Sample shape enters this relation, but not in a critical way, and we therefore ignore it except as it affects the value of the dc magnetic field at resonance, H_{res} . The relation is:

$$\lambda = \Delta H \gamma M_s / H_{res} . \quad (37)$$

From Bickford's²² data on line width and (37) we find $\lambda = 9 \times 10^8$ for Fe_3O_4 at room temperature. From the ferromagnetic resonance data reported elsewhere²⁴ on $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$, the material used in the present research, we find $\lambda \cong 10 \times 10^7$ at room temperature. Table I compares the room temperature values of λ obtained in the two ways on the two materials.

The differences between the domain wall experiments and the ferromagnetic resonance experiments lead to quite different behavior of λ in the two cases at low temperatures. These differences can be understood in terms of the frequency dependence of λ as given by an extension of the relaxation theory given in the third part of the theoretical discussion. A discussion of these relationships must await the detailed report on the ferromagnetic resonance results which is now in preparation, where such an extension will be given. As Table I shows, the room temperature values obtained in the two ways are of the same order of magnitude.

Let us now turn to a discussion of the relation between (31) and (33) and the data shown in Fig. 8. Qualitatively, of course, the $1/\tau$ factor in v/H_0 which (31) reveals, taken together with (32) explains most satisfactorily the sharp increase in viscous damping of the domain wall at low temperatures. Furthermore, it seems quite possible, although the author has not investigated it, that the higher order terms in (24) account for the nonlinearities in Figs. 9 and 10.

It should be mentioned that an increase in relaxation time at low temperatures which is consistent with (32) has been deduced by Bloembergen and Wang²⁶ and Healy²⁷ from ferromagnetic resonance data taken by them.

Quantitatively, we have inadequate data for a satisfactory comparison between Fig. 8 and (31), and the assumptions of the theory should perhaps be investigated further before any such comparison is taken

seriously. Nevertheless, plausible assumptions can be made which make an instructive comparison possible. $g_{1\infty}$ in (31) is equal to the difference in the variation of the anisotropy energy when measured adiabatically and when measured isothermally. At this stage of our knowledge, many assumptions which still retain the symmetry of the crystal are possible concerning the form of $g_{1\infty}(\theta)$. For our present purposes we will assume that it is given to within a constant by (34), but we must introduce a minus sign to take account of the fact that the rearranging electrons must have a *positive* anisotropy energy associated with them. Since we differentiate before substituting in (31), we do not need to subtract out the constant. The amplitude of $g_{1\infty}$ is not given by $|K_1|$ of course; we will call this amplitude $|K_r|$. When we introduce the minus sign into (34), and calculate the integral in (31), we find:

$$v = \frac{1}{\tau} \frac{4.0M_s}{|K_r| \sqrt{\frac{|K_1|}{A}}} H_0. \quad (38)$$

This relation points up the fact that the mechanism we are discussing here is characterized by two parameters, an amplitude factor $|K_r|$ and a time τ . Our experiment tells us nothing about $|K_r|$, so we will arbitrarily assume it is about $\frac{1}{2}$ of $|K_1|$ at room temperature, or 20000 ergs/cc. $|K_r|$ can be measured by comparing values of $|K_1|$ determined isothermally, say by measuring the torque on a disc, and values of $|K_1|$ determined adiabatically, say by ferromagnetic resonance experiments, but this has not been done as yet. To determine τ at 77°K, assume that the maximum frequency of rotation of dipoles in the wall is such that $\omega\tau \leq 0.1$ when the non-linearity shown in Fig. 9 first occurs (at $v = 150$ cm/sec). We calculate the thickness of the wall to be 4×10^{-5} cm from standard formulae (see Kittel's review article, Reference 14) and find $\tau \leq 0.5 \times 10^{-8}$. These values of τ and $|K_r|$, together with $|K_1|$ from Table II at 77°K and A as calculated above, give $v/H_0 \leq 40$ when inserted into (38). This is to be compared with a measured $v/(H - H_c)$ of 160 at 77°K. In view of the preliminary nature of (31), and the arbitrariness of some of our assumptions, this check is quite satisfactory. It would be naive to expect the theoretical result to be closer than an order of magnitude to the experimental one. Inspection of Fig. 8 suggests that the mechanism which gives the sharp increase in damping at 77°K is submerged at room temperature in the effects of other mechanisms, perhaps other electronic rearrangements, perhaps exchange effects of the sort recently suggested in metals by Rado.²⁸ Equation (39) confirms this expectation when we use $\tau \leq 10^{-12}$, as determined either

from ferromagnetic resonance data²⁴ or from the conductivity of Fe_3O_4 . The above method of calculating τ from domain wall data is less satisfactory at room temperature.

The idea of associating losses in ferrites such as this one with changes of order in the divalent and trivalent iron ions explains the fact that the damping observed in domain walls at room temperature in Fe_3O_4 ,⁷ where none of the divalent iron is replaced by nickel, is larger than that observed in $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$. Finally, as Wijn and van der Heide¹⁷ have pointed out, and as (31) and (33) show more analytically, this mechanism is a very satisfactory explanation of the sharp change in domain wall relaxation frequency observed by Galt, Matthias, and Remeika.²⁹

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to many friends and colleagues for assistance with various aspects of this research. The crystals from which these samples were cut were obtained through Dr. G. W. Clark from the Linde Air Products Co. H. J. Williams has been of considerable help in observing the domain patterns. Most of the work of cutting the samples from bulk crystal and preparing them was done by J. A. Andrus, Mrs. M. R. Tiner, and R. E. Enz. The hysteresis loops were taken in collaboration with P. P. Cioffi and F. J. Dempsey. The special pulser used in obtaining the data was designed by H. R. Moore and built by H. G. Hopper. The chemical analyses were made by H. E. Johnson, J. F. Jensen, and J. P. Wright. Enlightening discussions were had with J. F. Dillon, Jr., C. Herring, A. N. Holden, B. T. Matthias, W. T. Read, H. J. Williams, and A. M. Clogston, who derived independently a result somewhat similar to that in (31). Useful comments were made on the manuscript by R. M. Bozorth, S. Millman and P. W. Anderson.

REFERENCES

1. Rado, Wright and Emerson, *Phys. Rev.*, **80**, p. 273, 1950. Rado, Wright, Emerson and Ferris, *Phys. Rev.*, **88**, p. 909, 1952. G. T. Rado, *Revs. Mod. Phys.*, **25**, p. 81, 1953. Welch, Nicks, Fairweather and Roberts, *Phys. Rev.*, **77**, p. 403, 1950.
2. D. Polder and J. Smit, *Revs. Mod. Phys.*, **25**, p. 89, 1953. J. J. Went and H. P. J. Wijn, *Phys. Rev.*, **82**, p. 269, 1951. Wijn, Gevers and van der Burgt, *Revs. Mod. Phys.*, **25**, p. 91, 1953. H. P. J. Wijn and H. van der Heide, *Revs. Mod. Phys.*, **25**, p. 98, 1953. H. P. J. Wijn, Thesis, Leiden, 1953. Separaat 2092, N. V. Philips Gloeilampenfabrieken, Eindhoven, Holland.
3. H. J. Williams, *Phys. Rev.*, **52**, p. 747, 1937. H. J. Williams and W. Shockley, *Phys. Rev.*, **75**, p. 178, 1949. Williams, Bozorth and Shockley, *Phys. Rev.*,

75. p. 155, 1949. Williams, Shockley and Kittel, Phys. Rev., **80**, p. 1090, 1950. H. J. Williams, Bell Labs. Record **30**, p. 385, 1952.
- 3a. K. H. Stewart, Proc. Phys. Soc., **63A**, p. 761, 1950 and J. Phys. et Radium, **12**, p. 325, 1951.
4. K. J. Sixtus and L. Tonks, Phys. Rev., **42**, p. 419, 1932.
5. W. A. Yager and R. M. Bozorth, Phys. Rev., **72**, p. 80, 1947.
6. L. Landau and E. Lifshitz, Physik. Z. Sowjetunion, **8**, p. 153, 1935.
7. J. K. Galt, Phys. Rev., **85**, p. 664, 1952.
8. Galt, Andrus and Hopper, Revs. Mod. Phys., **25**, p. 93, 1953.
9. W. L. Bond, Phys. Rev., **78**, p. 646, 1950, Abstract I 10.
10. P. P. Cioffi, Phys. Rev., **67**, p. 200, 1945 and Rev. Sci. Instr., **21**, p. 624, 1950.
11. P. Weiss and R. Forrer, Ann. Phys., Series 10, **12**, p. 279, 1929. After the present work was completed, the author became aware of the extensive measurements of Pauthenet (Ann. Phys., Series 12, **7**, p. 710, 1952) on M_s in nickel ferrite and magnetite, among other materials. Interpolation between Pauthenet's values for these two materials suggests that the value of M_s we have used for $(\text{NiO})_{0.75}(\text{FeO})_{0.25}\text{Fe}_2\text{O}_3$ is about 6% too high at 445°K and 3% too low at 77°K, with intermediate errors between these temperatures and room temperature. Since, however, the accuracy of our data is not significantly better than this, and since our conclusions would be unaffected by any such changes, no correction has been made to our data.
12. R. Becker, J. Phys. et Radium, **12**, p. 332, 1951.
13. C. Kittel, Phys. Rev., **80**, p. 918, 1950; J. Phys. et Radium, **12**, p. 291, 1951.
14. C. Kittel, Revs. Mod. Phys., **21**, p. 541, 1949. See Equation 3.3.9. Since a wall perpendicular to the [100] direction in a crystal with a positive anisotropy energy constant is discussed in this reference, $g(\theta_0) = 0$ and is left out. The author is grateful to A. M. Clogston for pointing out the need for it in the present research. It was ignored in References 7 and 8 with the result that the values given for λ in those references were somewhat in error. Correct values are given in Table I.
15. E. J. W. Verwey and J. H. de Boer, Rec. Trav. Chim., **55**, p. 531, 1936. E. J. W. Verwey and P. W. Haaymann, Physica, **8**, p. 979, 1941. Verwey, Haaymann and Romeijn, J. Chem. Phys., **15**, p. 181, 1947.
16. M. E. Fine and N. T. Kenny, paper to be published.
17. H. P. J. Wijn and H. van der Heide, Revs. Mod. Phys., **25**, p. 99, 1953. H. P. J. Wijn, Thesis, Leiden, 1953. Separaat 2092, N. V. Philips Gloeilampenfabrieken, Eindhoven, Holland.
18. The author wishes to acknowledge a conversation with B. T. Matthias in which this mechanism was independently suggested in connection with the present research.
19. J. L. Snoek, New Developments in Ferromagnetic Materials, Elsevier, 1947.
20. It should be mentioned that Néel (J. Phys. et Radium **12**, p. 339, 1951; **13**, p. 249, 1952) has suggested that the after-effect losses discussed by Snoek¹⁹ in connection with the diffusion of carbon and oxygen in iron arise from an anisotropy relaxation. Néel's analysis, however, leads him to predict zero loss for large motions of a 180° domain wall, which is contrary to our experimental results. We suggest that he is led to an erroneous result because he allows the anisotropy energy itself to follow a relaxation of the form of Equation (18), whereas kinetic losses of this sort are due to the relaxation of one of two conjugate thermodynamical variables whose *product* is an energy. In our case these variables are the torque on the magnetization due to anisotropy, and the angle of rotation of the magnetization. It is this torque which satisfies the relaxation equation.
21. C. Herring and C. Kittel, Phys. Rev., **81**, p. 869, 1951. See Equation 5.
22. L. R. Bickford, Jr., Phys. Rev., **78**, p. 449, 1950.
23. C. Kittel, Phys. Rev., **76**, p. 743, 1949.

24. Galt, Yager and Merritt, Phys. Rev., **93**, p. 1119, 1954. A more detailed description of this work is in preparation.
25. Yager, Galt, Merritt and Wood, Phys. Rev., **80**, p. 744, 1950. See Equation (A-6). The λ of the present paper is equal to $\gamma \propto M_s$ in the notation of this reference.
26. N. Bloembergen and S. Wang, Phys. Rev., **93**, p. 72, 1954.
27. D. W. Healy, Jr., Phys. Rev., **86**, p. 1009, 1952.
28. G. T. Rado and J. R. Weertman, Phys. Rev., **94**, p. 1386, 1954.
29. Galt, Matthias and Remeika, Phys. Rev., **79**, p. 391, 1950.

Negative Impedance Telephone Repeaters

By J. L. MERRILL, JR., A. F. ROSE, and J. O. SMETHURST

(Manuscript received June 8, 1954)

In the exchange telephone plant, speech is transmitted largely at voice frequencies over a single pair of wires that carries conversation in both directions. Until recently, adequate transmission was assured through a suitable choice of coil loading and conductor size. The need for voice frequency amplifiers had long been recognized, but wide use of the conventional hybrid type of repeater with separate amplifiers for conversation in the two directions had not been economic. In 1948, however, a new type of repeater was introduced that used the same source of gain for transmission in both directions, and did not require costly filters. This repeater is known as the E1 repeater. It operates as a negative impedance in the line.

This paper describes the wide field of usefulness of two new types of negative impedance repeaters, the E2 and E3. Sufficient theory is given to show the advantages of using two types of negative impedance in a line, the series type and the shunt type, and how, by the addition of the shunt type to the series type, it is possible to insert gain without serious reactions on line impedance. The new equipment is described and maintenance features, together with methods of testing this relatively new type of repeater, are discussed.

APPLICATION IN THE BELL SYSTEM

INTRODUCTION

The application of the telephone repeater, the development of which made countrywide telephone service practicable, had been confined largely to toll plant from the year 1915 when the transcontinental line was first established, until a few years ago. About 1948 the negative impedance repeater¹ was developed and placed in production. This repeater operates on the principle of inserting negative resistance (and, if desired, negative inductance or capacitance) in series with the line, thus reducing the overall impedance and increasing the current in the line. This results in transmission gain in the same sense as that resulting from a repeater of the conventional type. This principle and the package

nature of the assembly resulted in a telephone repeater so low in cost and so simple in application and installation that it has found extensive use primarily in the local telephone plant.

In the period from 1948 to the present time, over 50,000 series-type negative impedance repeaters were manufactured and incorporated in the Bell System telephone plant. These repeaters have been used largely on intraexchange trunks and on trunks extending from the exchange areas to near-by smaller towns. Such installations have been very effective in improving the transmission on short haul calls and in many cases have also reduced trunk costs by permitting the use of smaller and cheaper conductors. They are usually operated at gains that reduce the nonrepeated trunk loss by more than half.

Negative impedance repeaters are especially suited to exchange trunk use, not only because of their simplicity and ease of maintenance, but also because unlike earlier types of repeaters, they preserve the dc continuity of the circuits on which they are installed. This latter feature means that they do not interfere appreciably with the signaling methods ordinarily used on such trunks. Also, they have the added advantage that, in the event of tube failure, the circuit still functions but with its loss substantially raised until the defective tube has been replaced.

APPLICATION OF SERIES-TYPE NEGATIVE IMPEDANCE REPEATERS

In a multioffice exchange area, trunks can generally be grouped in three classes. Largest in number are those known as interoffice trunks which extend directly from one local operating center to another local center in the same exchange operating area. In some cases direct trunks between two centers can not be justified and each office has trunks to a central location known as a tandem center where a through connection is made when required. This second class of trunks is known as "tandem" trunks. Frequently trunks of this type are provided to supplement the direct trunks and thus give the added advantage of alternate routing. A third type of trunk extends between the local office and the toll office and therefore always forms part of a toll connection to the local office. These are known as toll connecting trunks. In single office areas the toll connecting trunks to neighboring small towns are known as "tributary" trunks.

In some large multioffice exchange areas there are also trunks between the toll office and tandem centers which may be some distance from the toll office and through which local offices reach the toll office for long distance calls. These trunks are also called tandem trunks but are more

in the nature of intertoll links and are required to operate at very low loss similar to an intertoll circuit.²

The most extensive use of negative impedance repeaters has been in connection with the interoffice trunks. In this role they have been extremely useful and have contributed greatly to transmission improvement in the last few years. It would also be desirable to utilize this new and effective tool in reducing the losses of toll connecting and tandem trunks. There have been many cases of this type where improvement has been effected, but a very wide use has been prevented by their effect on return loss.

The insertion of a series negative impedance in an otherwise uniform loaded circuit introduces an impedance discontinuity which means that a substantial amount of energy is returned to the sending end of the circuit as "echo." As the introduction of negative impedance is the method of obtaining gain from the series repeater it follows that the magnitude of the impedance discontinuity and therefore the "echo," is proportional to the gain of the repeater. The effect of echo on ease of conversation depends both on its magnitude and on the amount of delay³ before the echo reaches the talker's ear. When the delay is small as in the case of an interoffice trunk a large amount can be tolerated, but in the case of a toll connecting trunk which becomes a part of a toll connection the delay may be large and only a small amount of echo can be permitted. This limits the amount of gain from a series repeater when it is used in toll connecting trunks.

As a result of this feature of the series negative impedance repeater, most cases requiring substantial gains in toll connecting or tributary trunks usually had to be handled by the older and more expensive hybrid-coil type repeater. With this latter repeater, it is practicable to introduce gain without serious reflection effects if proper attention is given to network design and to uniformity of construction of the line itself. However, the cost of the more complicated repeater was relatively high so that the new development described herein was undertaken to meet the indicated need, i.e., a repeater embodying the same desirable features of simplicity of design but still approaching in performance that of the hybrid-coil repeater in its effect on return loss.

The new repeater (E23) consists essentially of the earlier series repeater with the addition of a shunt negative impedance element. This combination has approximately as small an effect on return loss as the hybrid-coil repeater and will give about the same gain when used under similar line conditions. As a result, the field of use of the negative impedance type repeater will be greatly extended, especially in the toll

connecting plant where, from a practical standpoint, certain features of the hybrid-coil repeater are not needed. For example, more than 10 db gain is rarely required, and unequal gains in opposite directions would not ordinarily be useful. Besides, methods of signaling requiring a 4-wire split of the repeater are not utilized on exchange area trunks.

In addition to its important new characteristics, the series-shunt combination retains the desirable features of the earlier series repeater, i.e., it preserves dc continuity, it is simple in design and it is easy to install and maintain. If properly engineered and installed on circuits of reasonably uniform impedance, the new repeaters can be operated in tandem, do not have serious reactions on return loss, and are capable of reducing the losses of the trunks to about the same values as hybrid-type repeaters.

SCOPE OF APPLICATION OF THE NEW REPEATER

With the reduction of the return loss reaction of the earlier form of negative impedance repeater, the extent of application of the new E23 repeater becomes largely a matter of economics. In the telephone trunk plant, there are generally three gauges of conductors available for trunk use, namely, 19, 22 and 24. The larger gauges, utilizing more copper and requiring more conduit space, naturally cost more than the smaller

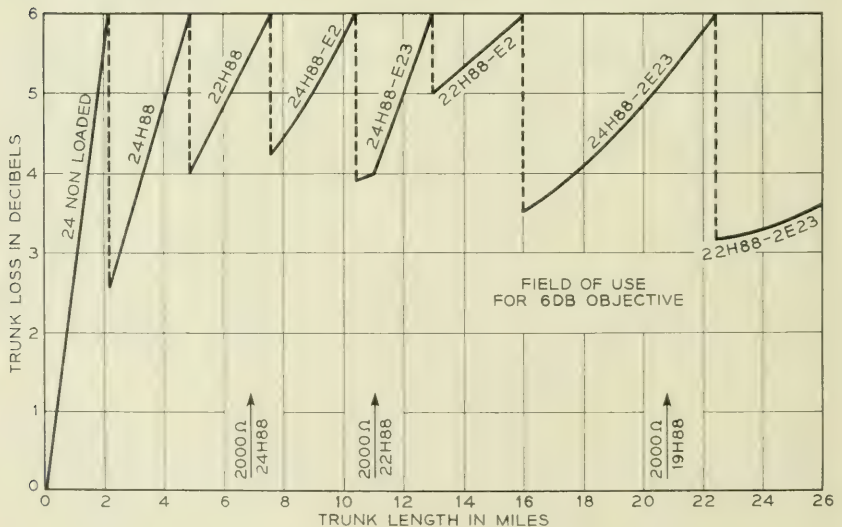


Fig. 1 — Illustrative field of use; interoffice trunks.

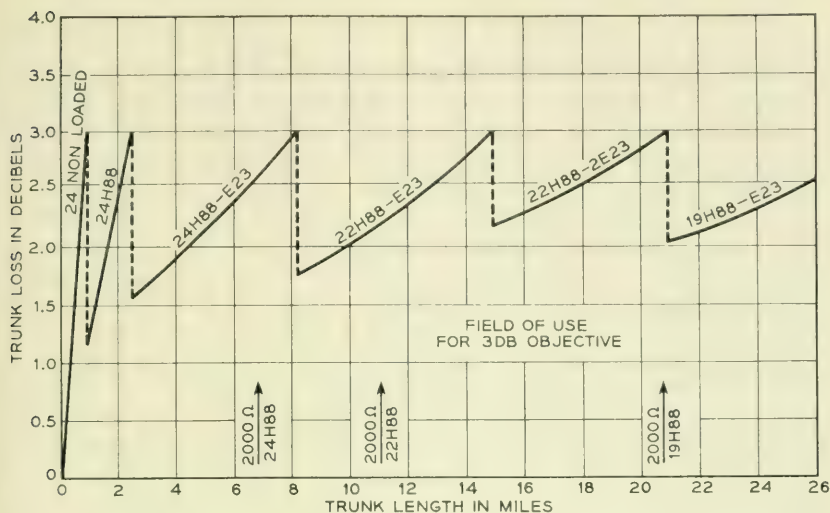


Fig. 2 — Illustrative field of use; toll connecting or tandem trunks.

gauge. In general, it is the practice to select the smallest gauge which will give the required transmission loss. For short distances, the differential in cost between trunks of different gauges is not large but as trunk length increases it will be found that a smaller gauge trunk with a repeater will cost less than one of larger gauge without a repeater. For example, at 10 miles a 22-gauge circuit with a repeater will cost substantially less than a 19-gauge circuit with no repeater. Also, as transmission objectives are improved, there may be cases where a smaller gauge trunk with two repeaters will be cheaper than a larger gauge with one repeater. Furthermore, in cases where the costs are about equal or even where the larger gauge is slightly cheaper, it will be found that lower losses can be obtained with repeatered circuits and the engineering choice would accordingly favor the smaller gauge.

Construction costs differ considerably depending on local conditions; hence the differential cost between the various gauges of conductors will differ correspondingly. However, for illustrative purposes Figs. 1 and 2 have been prepared which utilize average Bell System costs for both outside plant and telephone repeaters. In the case of the E23 repeaters, the experience so far has been somewhat limited so there is a greater degree of uncertainty as to the actual costs than for the older type repeater or the outside plant construction.

Fig. 1 shows the field of use for various conductor gauges when used in interoffice trunks where an over-all transmission objective of 6 db for

the trunks has been assumed. In considering this chart it will be seen that non-loaded 24-gauge conductors can be utilized for the shorter distances up to about two miles. The next step is to apply loading, which is indicated in the chart by the symbol H88 meaning 88 millihenry coils at 6,000 foot spacing. After reaching the limit of about 7 miles without repeaters on 22-gauge, the differential between 22- and 24-gauge is more than enough to pay for a repeater, the simple series repeater (E2) being used for distances up to 10 miles and then the improved E23 repeater extending the use of 24-gauge to about 14 miles. Beyond this point the most economical combination is indicated.

The general effect of the introduction of the negative impedance repeater is to shift the average gauge distribution so that more small gauge cable can be economically utilized, accompanied in most cases by improved transmission. It will be noted from the figure that no 19-gauge cable will need to be added in the future to care for interoffice trunks up to distances as great as 25 miles between offices. Of course, in cases where 19-gauge is already available in plant it can be used to advantage despite the fact that for new construction a smaller gauge with the negative impedance repeaters would be cheaper.

It has also been assumed in making up Fig. 1 that supervision or pulsing requirements will not limit the use of the smaller gauges. In a specific case where some of the older type central offices are involved, signaling may have an important bearing and substantially distort the economic ranges indicated on the chart.

Fig. 2 illustrates the field of use of toll connecting or tandem trunks. As a toll connecting trunk forms part of a multilink connection, and since there are always at least two (and often more) trunks in series on connections involving these trunks, the Bell System companies find it economical to plan for a maximum loss of 3 db for this type of trunk. Likewise for the tandem trunks, where two in series may be used in place of an interoffice trunk, 3 db is considered a reasonable objective.

It will be noted again by referring to this chart that 19-gauge has very little future field of use and in all cases the new "series-shunt" repeater will be utilized rather than the earlier series type because of return loss considerations and also because of the greater gain required to reduce the trunk losses to the desired values. In the tandem and toll connecting case, signaling may again be a distorting factor though not to so great an extent as in the direct interoffice trunk case. To this extent, however, the curves are theoretical, as they have been made up without regard to this limitation which may apply in a few practical cases.

SPECIFIC APPLICATIONS

Many installations of the new repeaters have been engineered for completion in 1954. In all cases economic studies were made and results of these studies broadly confirmed the indications of the two charts discussed above. To bring out more clearly the effectiveness of the E23 repeater, it may be worth while to consider a few specific cases involving installations of the new repeaters being made this year.

The first case shown on Fig. 3 is in the area of the Pacific Telephone and Telegraph Company. The figure shows the two alternatives that were considered for providing additional tandem trunks between San Francisco proper and the East Bay Area, needed this year because of an extension of customer toll dialing arrangements.

The engineering study for this project was somewhat more complicated than would ordinarily be the case because two possible routes of unequal length and with slightly different amounts of submarine cable were involved. The present route which touches Yerba Buena Island is subject to some hazard from dragging ship anchors but the circuit relief would have been cheaper here than on the other route shown, were it not for the new repeaters. The second route is considerably less hazardous and, in addition, is sufficiently removed from the present one to provide increased reliability under disaster conditions. Without the repeaters, however, the second route would have been somewhat impracticable, since it does not allow easy installation and access to required loading points between the two shore lines. However, with two of the new repeaters on each pair of conductors, nonloaded 22-gauge cable gives substantially the same effective transmission loss as loaded 19-gauge conductors on the shorter route and over a future period will involve less annual cost per circuit. Furthermore, the use of 22-gauge cable permits more than twice as many pairs to be included in the same size sheath in the expensive submarine section.

The initial installation of the new type repeaters on this cable will total about 1,300, divided between the Main Office in San Francisco and the Main Office on the East Bay side. Individual repeater gains are 5 to 7 db at 1,000 cycles but the repeater gain characteristic is shaped to offset the increasing cable losses at the higher frequencies so that the over-all circuit has relatively uniform transmission in the voice range.

The second case is one where the need for the repeaters results from the complex toll switching system at Chicago, Illinois. Here it has been found desirable, because of the great volume of toll traffic, to have several toll offices scattered throughout the city and suburbs. In general, toll calls coming into these offices from other cities are completed over toll

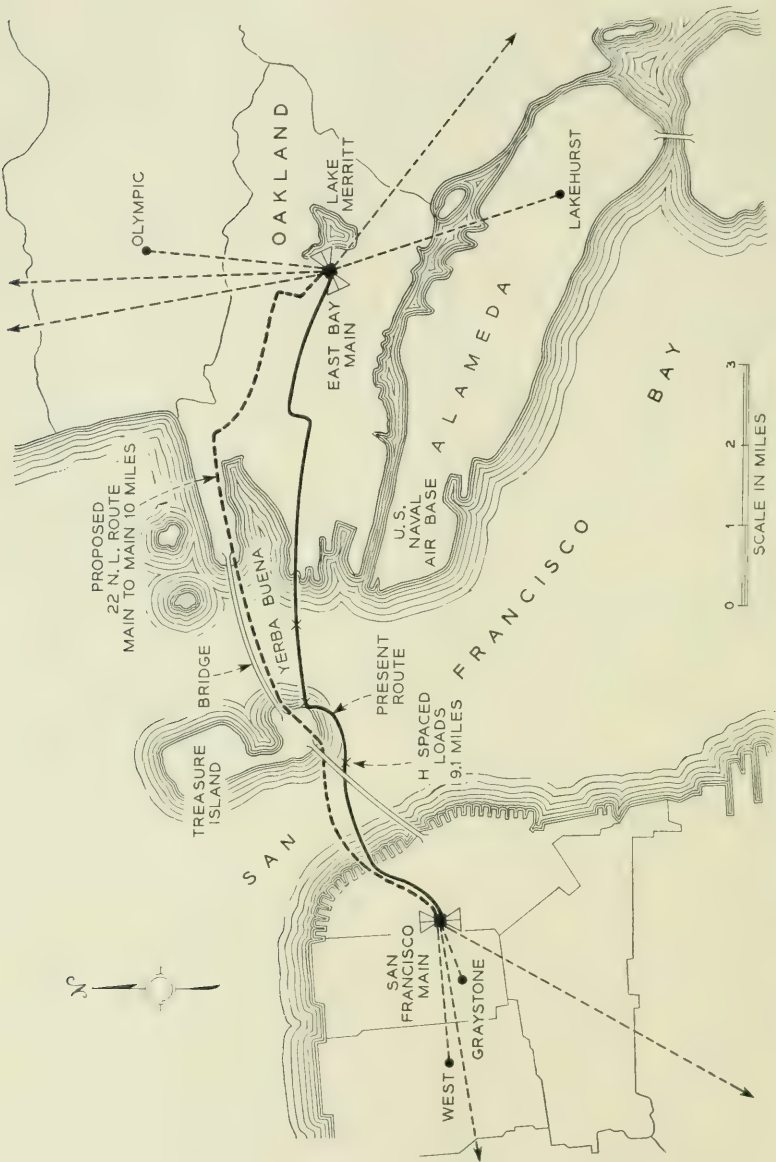


Fig. 3 — Typical E23 repeater application, San Francisco — East Bay tandem trunks.

connecting trunks direct to the called subscribers' central offices. It is not economical, however, to provide enough such trunks to handle peak loads. When all of these direct trunks are busy, an incoming toll call is switched to the subscriber's central office via a tandem office serving his general area. Since the losses of trunks between the tandem and local offices are of the same order as those of the direct trunks from toll to local offices, it would be desirable to operate the trunks between the toll and tandem offices at losses close to 0 db if this were practicable. In this way the same transmission objectives would be met on both routings and no contrast in transmission would be evident to the same subscribers on calls completed over the different routes at different times. Fig. 4 illustrates a specific case of 36 trunks between toll office No. 3

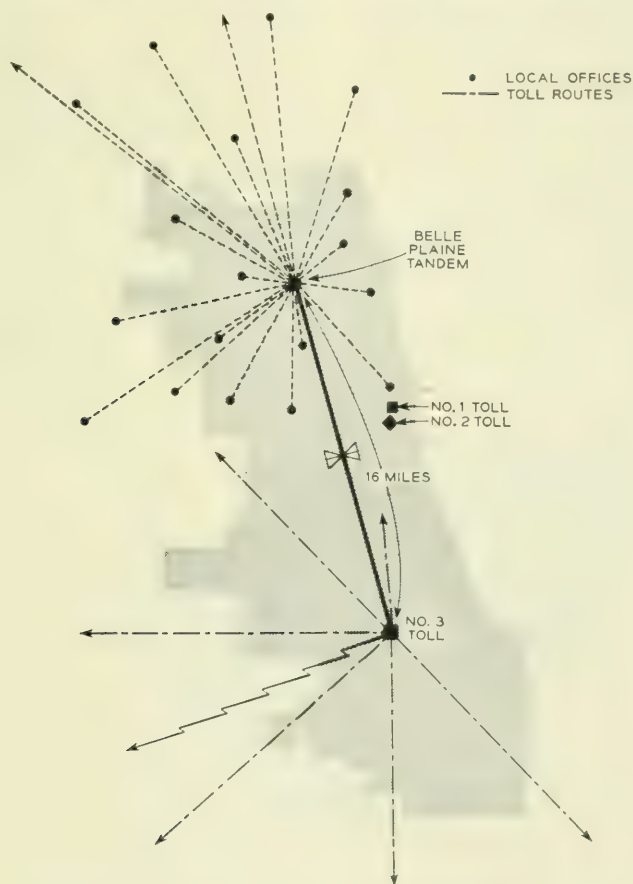


Fig. 4 — Typical E23 repeater application; Chicago toll tandem trunks.

and Belle Plaine tandem, a distance of about 16 miles. Without repeaters these trunks would have been operated at a loss of about 10 db but with the repeaters they are operated at about 2 db. The only alternative to using E repeaters in this specific case would have been to use the more expensive hybrid-type repeaters.

In some cases similar to this one, it was found practicable to temporize by installing the series repeater first and operating it at limited gain until the shunt element became available. Later, when the production of the shunt element was started, the additional units were added and full advantage of the new series-shunt design utilized in reducing the equivalent of the trunks to the lowest value permitted by echo return loss considerations.

The third example shown on Fig. 5 is in the city of Pittsburgh between Churchill tandem and the town of New Kensington, approximately 17 miles. Here the transmission on existing 19-gauge loaded cable without a repeater would have been about 8 db. The repeaters reduce this figure to between 2 and 3 db which should be satisfactory in this case. It should be noted, however, that 22-gauge with two repeaters would also provide a low enough equivalent, and should it become necessary to supplement the present cable at some future date, there will be an opportunity for further savings from the E23 repeaters.

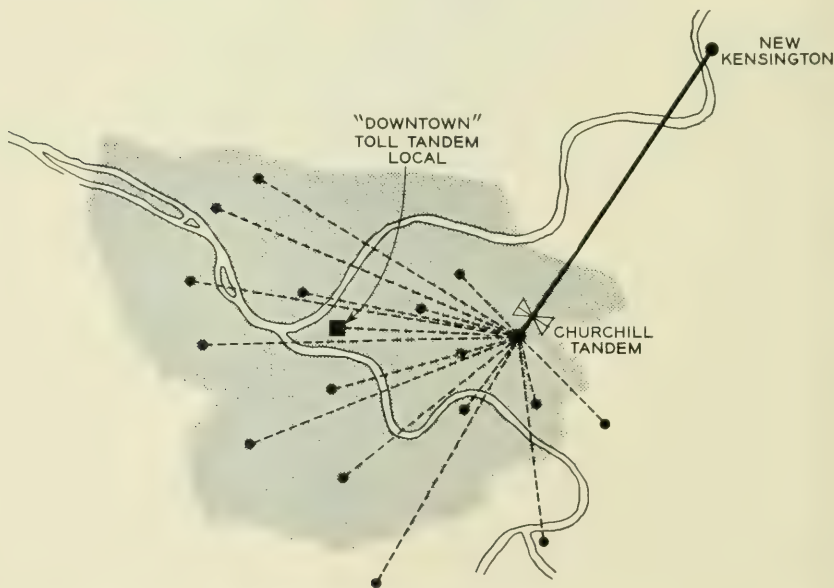


Fig. 5 — Typical E23 repeater application; Pittsburgh toll connecting trunks.

THEORY

NEGATIVE IMPEDANCE CONCEPT

Both the E2 and E3 repeater elements contain an amplifier having multiple feedback paths. The operation of an amplifier circuit of this type can be explained by classical feedback theory. However, experience with the E1 repeater over the past four years has shown the value of using a negative impedance concept in engineering such a device. Hence, in the explanation of operation given here, the repeater units will be treated simply as two-terminal networks which have negative impedance inputs over the frequency band of interest. The effect of introducing these impedances into telephone circuits can then be computed by the same simple network theory used to determine the effects of passive impedances.

THE E2 REPEATER UNIT

The E2 repeater is essentially a two-terminal network the impedance of which has a magnitude $|Z|$ and a negative phase angle that can vary with increasing frequency from minus 90 degrees, or less, through minus 180 degrees to at least minus 270 degrees. This type of negative impedance is shown in the diagram of Fig. 6(a). It has been known for many years as the series type because it could be produced by connecting the output of an amplifier back in series with its input. More recently it has come to be known as an open circuit stable, negative

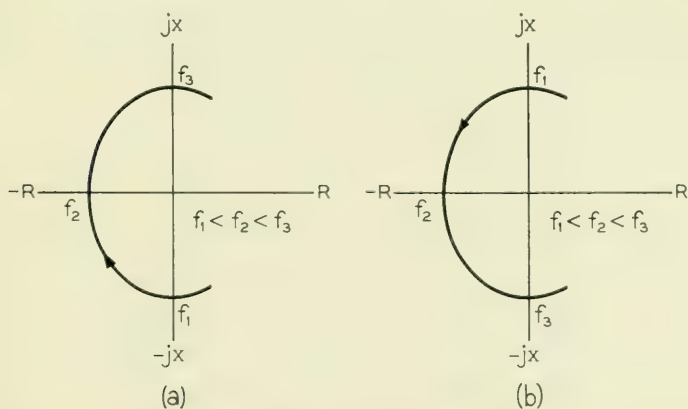


Fig. 6 — The two types of negative impedance: (a) Open Circuit Stable and (b) Short Circuit Stable.

impedance, because it will not oscillate when its two terminals are open circuited.

THE E3 REPEATER UNIT

The E3 repeater is essentially a two-terminal network the impedance of which has a magnitude $|Z|$ and a positive phase angle that can vary with increasing frequency from plus 90 degrees, or less, through plus 180 degrees to at least plus 270 degrees. This type of negative impedance is shown in Fig. 6(b). It has been known as the shunt type because it could be produced by connecting the output of an amplifier back in shunt with its input. In more recent years it has become known as the short circuit stable type because it will not oscillate when its two terminals are short circuited.

THE NEGATIVE IMPEDANCE CONVERTER

The amplifier circuits of both of these repeater units perform the same function: that of a negative impedance converter. The operation of such converters is illustrated in Fig. 7.

Fig. 7(a) shows the converter as a four-terminal network having a ratio of transformation k and a shift of phase through a negative angle of approximately 180 degrees over the operating band of frequencies. If, as shown, an impedance Z_N is connected to terminals 3 and 4 then the impedance seen at terminals 1 and 2 will be the impedance Z_N multiplied by the ratio k and shifted in phase through a negative angle of 180 degrees. This impedance will (over the frequency range of zero to infinity) fulfill the definition given for the impedance presented by the E2 repeater. Hence, Fig. 7(a) can represent the operation of the E2 repeater.

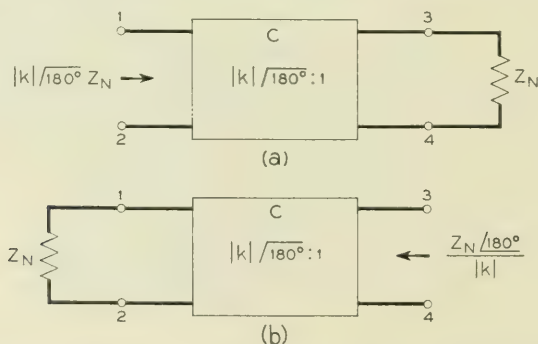


Fig. 7 — The Negative Impedance Converter: (a) E2, open circuit stable and (b) E3, short circuit stable.

Fig. 7(b) shows the same converter, but here the impedance Z_N is connected to terminals 1 and 2. The impedance seen at terminals 3 and 4 (at least over the frequency band of interest) will be Z_N divided by k and shifted in phase through a positive angle of approximately 180 degrees. This impedance will (if frequencies from zero to infinity are considered) fulfill the definition given above for the E3 repeater impedance. Thus Fig. 7(b) can represent the operation of the E3 repeater.

From Fig. 7, it is apparent that the same converter circuit could have been used for both the E2 and the E3 repeaters. For practical reasons it was not. However, the ratio k and the phase shift in both the converter of the E2 and that of the E3 were made approximately the same.

OPERATION IN TRANSMISSION LINES

Within limitations, the E2 repeater can be represented by a negative impedance, $-Z$, and the E3 repeater can be represented by a negative admittance, $-Y$. With a negative impedance and a negative admittance available, losses of transmission lines can be reduced in the manner illustrated in Fig. 8. The transmission line is represented by two networks as shown in Fig. 8(a). One of these (Network A) is in the form of a T network the series arms of which are represented by impedances Z ; and the shunt arm, by an admittance Y . This network has a propagation constant $\alpha_1 + j\beta_1$. The attenuation α_1 represents the major portion of the line attenuation, and the phase shift β_1 is that just sufficient to make Network A realizable physically. This representation is necessary because Network A has image impedances each equal to the characteristic impedance (Z_0) of the line. If the characteristic impedance of the line were a pure resistance, then the phase shift through this network could be zero and β_1 could be zero. But the characteristic impedances of actual lines are not pure resistance; thus the phase shift β_1 must be included in Network A. The other network (Network B) is shown as a box. It has a propagation constant $\alpha_2 + j\beta_2$. Here β_2 represents the remaining phase shift in the transmission line and α_2 is an attenuation just sufficient to make Network B physically realizable in view of the image impedances which are both equal to Z_0 , the characteristic impedance of the line. Fig. 8(b) shows the addition to this line of a repeater consisting of a T network made up of negative impedances $-Z$ in the series arms and a negative admittance $-Y$ in the shunt arm. The arm $-Z$ of the repeater adjacent to the line cancels Z of the line. The two admittances $-Y$ and Y cancel and the other series arms $-Z$ and Z also cancel. The result, as shown in Fig. 8(c), is that only the attenuation and phase shift of Network B remain.

In practice the amount of attenuation (α_1) which can be canceled by the repeater depends on the uniformity, the loss and the type of line. The permissible magnitude is computed by conventional methods which are beyond the scope of this paper.

Fig. 8 has shown how the combination of a series and a shunt repeater can annul a large part of the attenuation of a telephone line. Much may be accomplished also by a series negative impedance alone as illustrated in Fig. 9. In Fig. 9(a) a transmission line is again represented by two networks A and B. However, Network A now is shown as an L configuration having a series arm Z and a shunt admittance Y together with an ideal transformer of ratio $1:K$ where K exceeds unity. Network A of Fig. 9 is equivalent to Network A of Fig. 8, and Network B of Fig. 9 is the same as Network B of Fig. 8. In Fig. 9(b) the addition to this transmission line of a single $-Z$ such as the E2 repeater is shown. This negative impedance $-Z$ cancels the series impedance Z of Network A. The result is shown in Fig. 9(c).

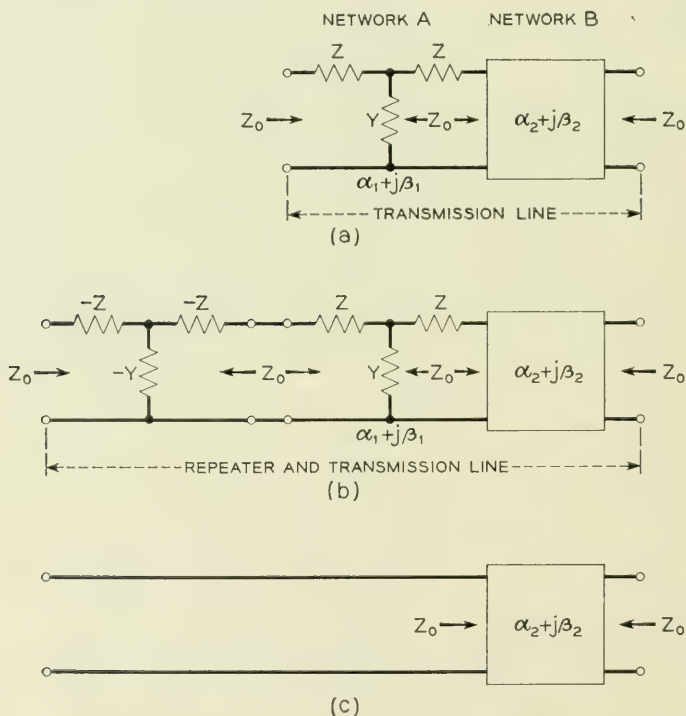


Fig. 8 Operation of the "T" in a transmission line: (a) transmission line, (b) repeater and transmission line, and (c) result of addition of repeater.

Thus Fig. 9 shows that when a single $-Z$ is added in series with the conductors of a transmission line the attenuation is reduced, but the equivalent of a shunt conductance Y together with an impedance transformation $1:K$ could be corrected by means of a transformer if it were not for the shunt element Y . Thus an impedance irregularity is introduced, and this irregularity reflects power back toward the source. When the reflected power becomes a significant part of the total power passing through the line, transmission is unsatisfactory. Echo is excessive and therefore the use of the single series repeater is limited.

THE BRIDGED T STRUCTURE

The discussion of the T repeater, illustrated in Fig. 8, was based on the use of two series negative impedances and a shunt negative admittance. It is perfectly possible, and more economical, to obtain the same effect by using a single series impedance in a bridged T structure and this

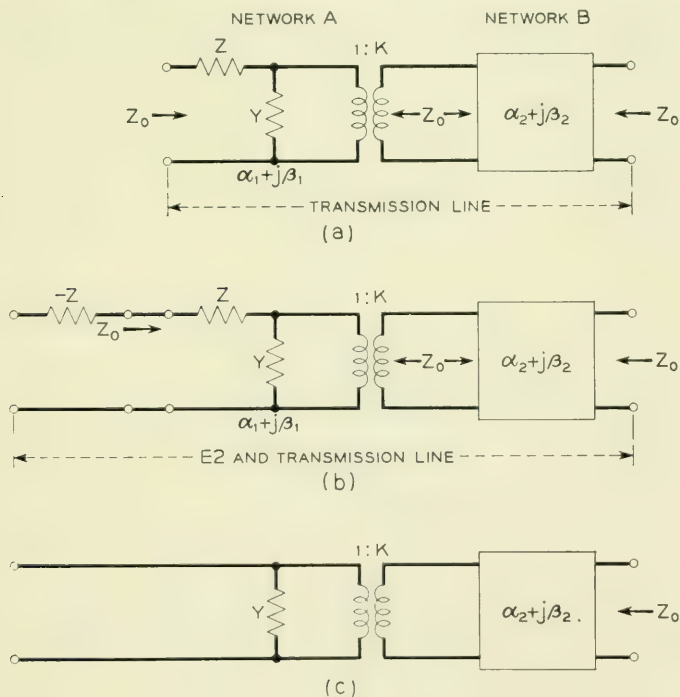


Fig. 9 — Operation of $-Z$ in a transmission line: (a) transmission line, (b) repeater and transmission line, and (c) result of adding repeater.

is, in fact, what is done with the E23 combination repeater. Fig. 10 shows how this is accomplished by using a center tapped line coil for the E2 repeater with the E3 connected as a shunt element to the midpoints of the coil. If the coil is considered to be ideal this arrangement is equivalent to the configuration shown in Fig. 11 in which the coil provides the basic bridge structure: the E2 repeater is the series arm; and the E3, the shunt arm. Incidentally this arrangement of E2 and E3 repeaters is similar to G. Crisson's twin 21-type repeater.⁴

For a bridged-T structure, the image impedance equals the square root of the product of the series and shunt arms and the attenuation (in db) is as indicated on Fig. 11.

If a network is to be inserted in an electrically long line without introducing an irregularity, its image impedance must match the characteristic impedance of the line. This would be the case for the bridged T network if Z_A were set equal to NZ_0 and Z_B were set equal to Z_0 divided by N . Then the square root of the product of Z_A and Z_B would be Z_0 .

A network made up of negative impedances is designed to match a line in the same way and Fig. 12 is a representation of such a structure. Here, as a matter of convenience, the shunt arm is shown as an impedance

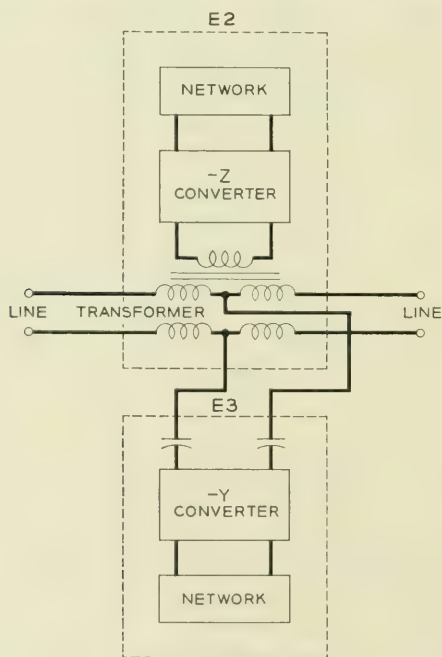


Fig. 10 — The bridged T repeater E23.

with a $+180^\circ$ phase shift instead of an admittance as used previously. The letter N designates a numeric or proportionality constant. It will be observed that the product of the shunt and series impedances is a real or positive impedance and hence the image impedance is a positive impedance, Z_0 . The gain is determined entirely by the value of N . Thus if the characteristic impedance of a transmission line is known, together with the gain that the line can support without risk of oscillation, then N is known and the repeater network can be adjusted to give the required gain.

The advantage of the bridged T as compared to a single series negative impedance such as the E2 can be demonstrated by comparing the relative transmission gains obtainable from the two arrangements. Fig. 13(b) shows the insertion loss of a single impedance Z_A connected in series with a transmission line having a characteristic impedance Z_0 . If Z_A is a negative impedance such as that produced by the E2 repeater then the repeater gain becomes a function of N as shown in Fig. 13(c). If N equals 2 the gain is infinite and the system will oscillate. Thus N must always be less than 2 where Z_A is a negative impedance of the series or open circuit stable type. Practically, the impedance of the transmission line is not a constant Z_0 but varies with termination, line construction and temperature. Thus N should be decreased until the negative impedance is always less than the sum of the two line impedances in series with it taking into account all possible variations in these impedances.

The same limitations on N apply to the bridged T repeater of Fig. 12

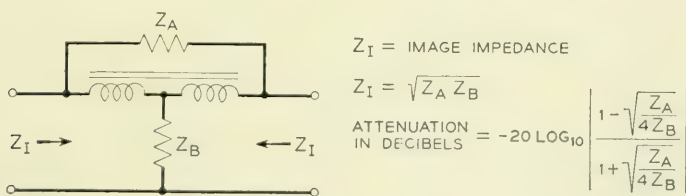


Fig. 11 — Schematic of the bridged T network.

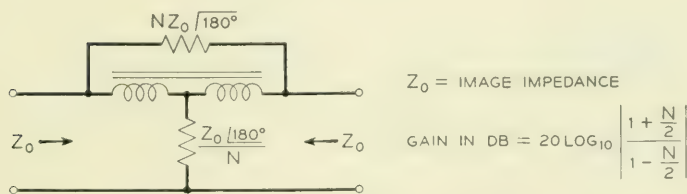


Fig. 12 — Schematic of the bridged T repeater.

as apply to the single series repeater. If N is 2 the gain is infinite and the circuit will sing. This similarity goes further. Assume that a single negative impedance Z_A equal to $NZ_0/180^\circ$ is inserted in series in an electrically long line and N is adjusted for stability. If this series element is removed and the bridged T of Fig. 12 is inserted in the same place and adjusted by changing N until the system is stable it will be found that N will have the same value in the bridged T structure as it had for the single series negative impedance.

Thus, if N is the same in the case of the bridged T as in the single series impedance, the gain advantage can be obtained by comparing formulas on Figs. 12 and 13 from which it can be seen that the gain advantage of the bridged T is equal in db to $20 \log_{10} [1 + (N/2)]$. If a single series repeater can be used in a line to give an insertion gain of 6 db ($N = 1$) then a bridged T can be used to provide $20 \log_{10} (1 + 0.5)$ or 3.5 db additional. Thus, in this case the series repeater gives 6 db gain as compared to $6 + 3.5$ or 9.5 db for the bridged T. These gains are theoretical; in actual lines with simply constructed repeaters the comparison may not be quite so favorable to the bridged T.

THE NEGATIVE IMPEDANCE CONVERTER

So far the discussion of the E2 and E3 repeaters has been in terms of a "black box" which translates a positive impedance into a negative

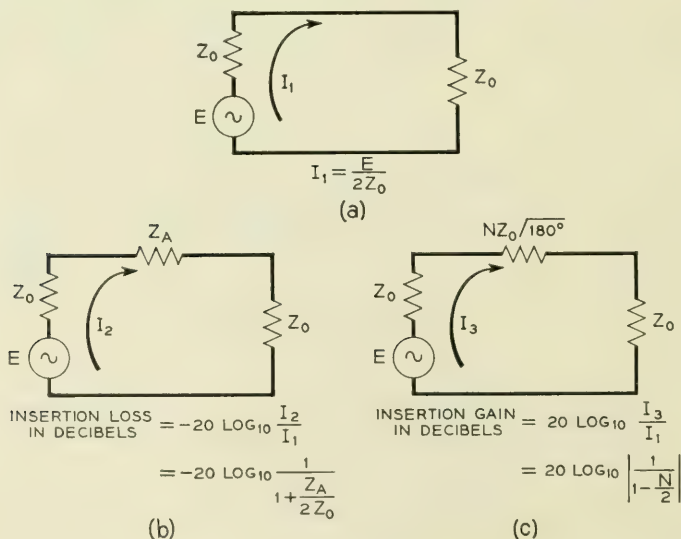


Fig. 13 — Insertion gain of the E2 repeater.

impedance through a multiplying and phase shift operation. It will be interesting to examine these boxes to see what factors determine their characteristics.

THE E2 CONVERTER

The E2 negative impedance converter is the same as the E1. As discussed elsewhere⁵ it can be represented schematically as in Fig. 14(a) and also in terms of the equivalent circuit of Fig. 14(b) if the coils are assumed to be ideal. The converter performs much like a transformer. An impedance seen through it is not only transformed in magnitude by the ratio of $|1 - \mu\beta|$ to $|1 + \mu|$, it is also modified by the phase shift of this factor which over the operating band of frequencies approximates 180 degrees. The symbol μ stands for the voltage gain of the electron tube and β is the ratio of 1 to $1 + (j\omega CR)$. If both C and R are large β approaches unity in magnitude and the ratio of conversion approaches $1 - \mu$ to $1 + \mu$. If μ is large compared to unity then this conversion ratio approaches -1 . This ratio of -1 is approximately realized in the E2 converter, and therefore the conversion ratio is not changed appreciably by small variations in μ .

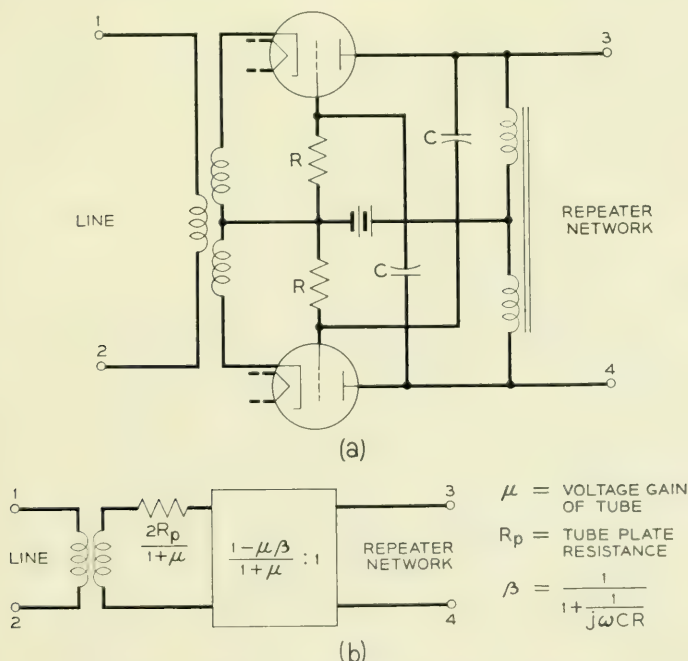


Fig. 14 — E2 Converter; (a) schematic and (b) equivalent circuit.

In addition to the transformation term there is also, as shown in Fig. 14(b), a series term, $2R_p$ divided by $1 + \mu$. Here R_p is the plate resistance of the tube, and μ is the voltage gain as mentioned before. The factor 2 results from the use of two tubes in push pull. If μ is large compared to unity then this series term becomes approximately $2R_p/\mu$. It is entirely dependent upon the characteristics of the electron tube. As the characteristics change from tube to tube with manufacturing variation or in the same tube over a period of time or with variation in battery supply potential, the term $2R_p/(1 + \mu)$ will change accordingly. Percentage-wise this change may be large. This is the largest source of variation in the E2 converter. It can be minimized by operating the converter between impedances much larger in magnitude than $2R_p/(1 + \mu)$ so that variations in this term have relatively small effect. This has been done in the E2 repeater by stepping up the impedance of the transmission line by about 1:9 by means of the transformer shown in Fig. 14.

THE E3 CONVERTER

Theoretically, the same converter used for the E2 and shown in Fig. 14 could have been used for the E3. Instead of connecting the line to

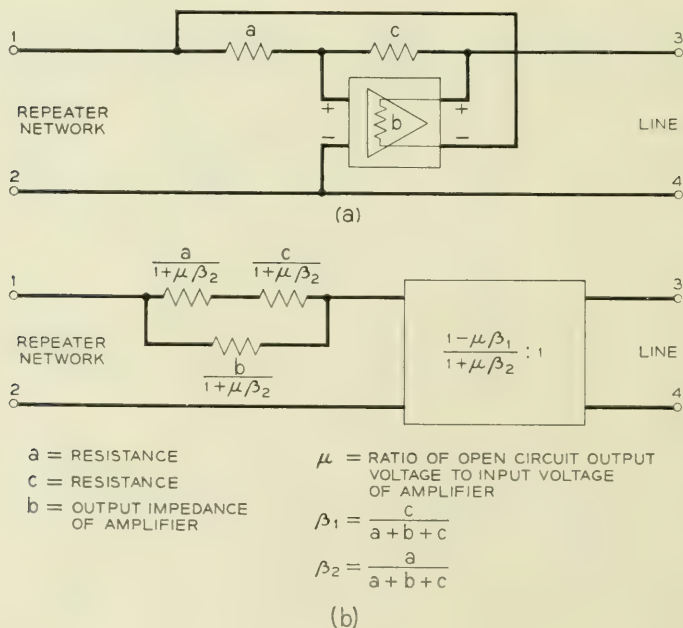


Fig. 15 — E3 Converter; (a) schematic and (b) equivalent circuit.

the converter through terminals 1 and 2, terminals 3 and 4 would be used. However, because the E3 must be designed for connection across a transmission line a coil or transformer input is not practical since the coil would shunt the line at low frequencies and introduce excessive loss to dial pulsing and 20 cps ringing. Without a coil to step up the impedance of the line, variations in $2R_p(1 + \mu)$ with standard triodes are too large to be neglected. For this reason another converter circuit was designed for the E3 repeater.

This circuit is shown in schematic form in Fig. 15(a). It consists of two resistances, a and c , respectively, and an amplifier poled according to the plus and minus designation on Fig. 15(a). The output impedance of the amplifier has been designated as b . If the input impedance of this amplifier is high compared to other circuit impedances, Fig. 15(a) can also be represented by the equivalent circuit of Fig. 15(b). Here is a conversion factor similar to that in the E2 converter and also a series impedance. The factor μ is the ratio of the open circuit output voltage to the input voltage of the amplifier. In the E3 converter this voltage ratio μ is quite high because the amplifier is a two-stage arrangement. In the design of the E3 both β_1 and β_2 are approximately one half. Thus $\mu\beta_1$ and $\mu\beta_2$ are both large compared to unity so that the conversion ratio $(1 - \mu\beta_1)/(1 + \mu\beta_2)$ is approximately unity and relatively independent of variations in μ . Furthermore, because $\mu\beta_2$ is large compared to b , the series term in the converter circuit is relatively small and variations in this term have little effect on the operation of the converter.

CIRCUIT DESCRIPTION

THE E2 TELEPHONE REPEATER

The circuit function of the E2 telephone repeater can be divided into two parts: the electron tube (negative impedance) converter; and the adjustable two-terminal network associated with the converter.

In order to reduce the effect of variations in the electron tubes to negligible proportions, and at the same time to operate the tubes with load impedances that will permit optimum energy transfer from tube to connected circuit, the impedance of the telephone line is stepped up by means of the input transformer. To insure adequate balance for use in the telephone lines, the low voltage side of the transformer is divided into two equal, balanced windings. Each winding is center-tapped and connected in series with a line conductor. The circuit of the E2 repeater is shown in Fig. 16.

In practice, it is advantageous to limit the conversion bandwidth so

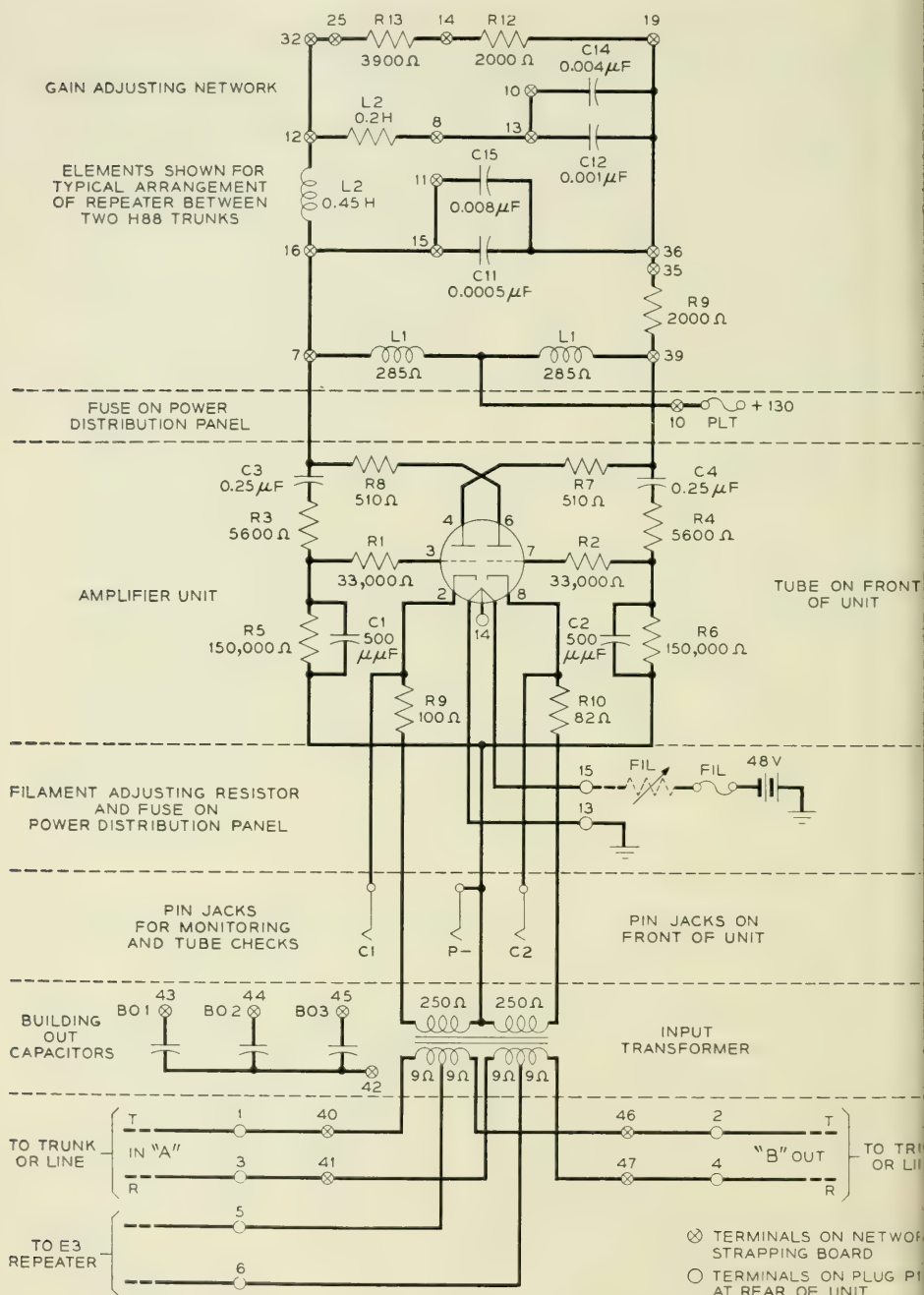


Fig. 16 — Circuit of E2 repeater.

that the line and network impedances do not have to be controlled over an extended frequency band. The conversion gain is limited by reducing β to a small value at low frequencies by capacitors C_3 and C_4 of Fig. 16 in the plate-to-grid coupling network, and at high frequencies by the small capacitors C_1 and C_2 across the grid resistors R_5 and R_6 respectively.

The conversion ratio is affected by small losses inherent in the electron tube and transformer. These are balanced out by a fixed resistor R_9 connected in series with the gain adjusting network to increase the amount of positive feedback.

The final negative series impedance presented to the line is equal to approximately $-0.1Z_N$ over the frequency band of 300 to 3,500 cycles. The impedance Z_N is determined by the configuration of the gain-adjusting network comprising several inductors, capacitors, and resistors. These components may be arranged in any form to obtain the desired negative impedance, which in turn, introduces the gain and frequency shaping characteristic desired for each type of line facility.

The E2 repeater employs a Western Electric 407A twin-triode electron tube of the 9-pin miniature type. The tube heater circuit can be operated from 24- or 48-volt office battery. The heater current is 100 milliamperes for 20-volt operation, 50 milliamperes for 40 volt operation and the plate current is 11 milliamperes.

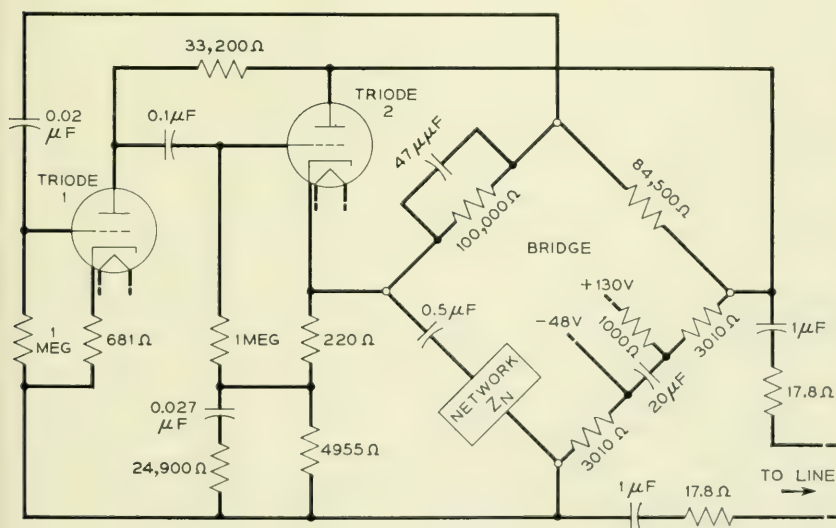


Fig. 17 — Schematic circuit of E3 repeater.

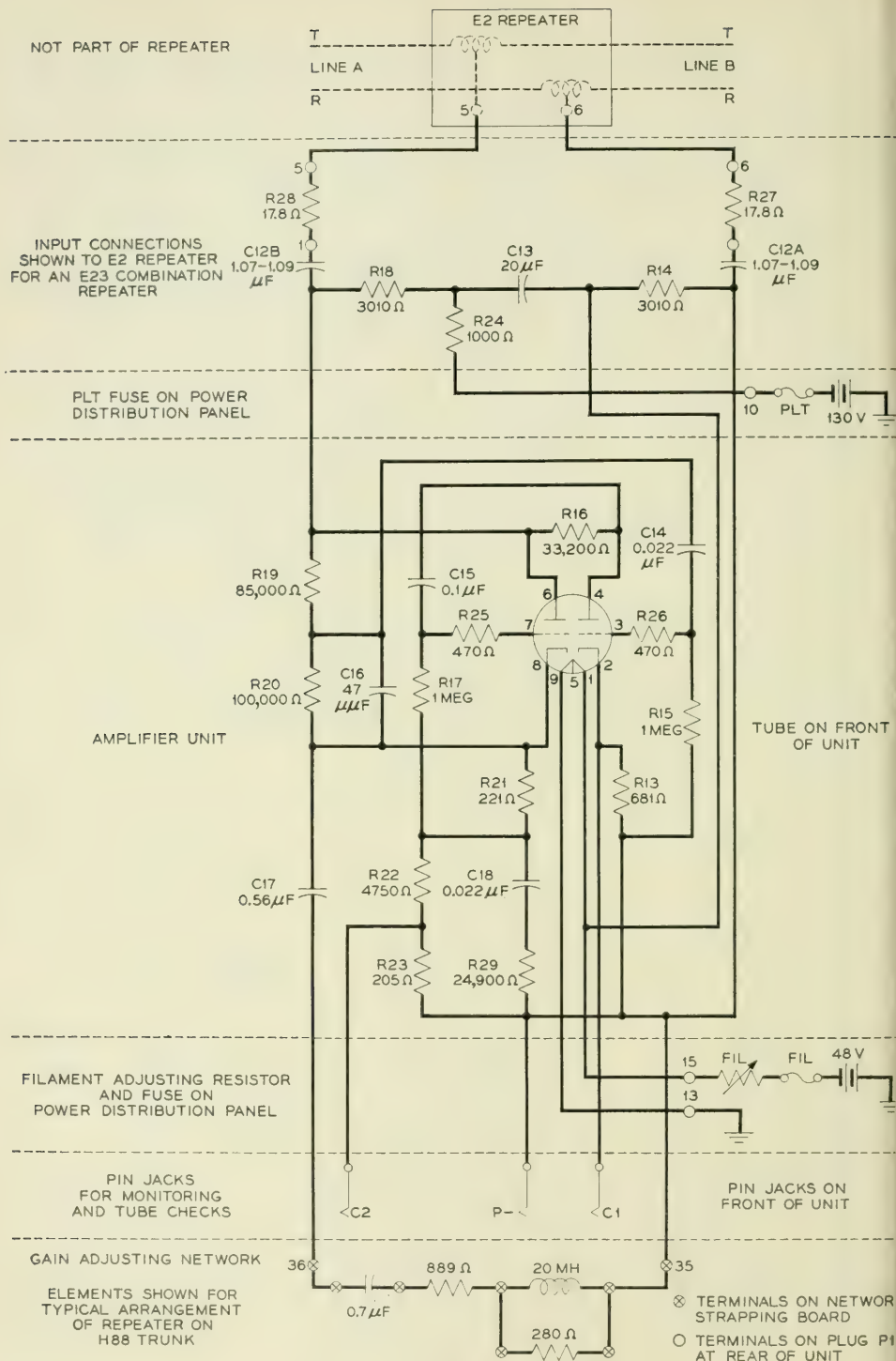


Fig. 18 — Circuit of E3 repeater.

THE E3 TELEPHONE REPEATER

The E3 repeater is a two-terminal high impedance device designed to be bridged across the line in contrast with the E2 repeater which is a low impedance series connected device. Schematically shown in Fig. 17, its circuit is considerably different from that of the E2. The bridged telephone line and an adjustable network form two arms of a bridge connected to output and input circuits of a two-stage amplifier.

The bridge is so connected that a fairly large proportion of the amplifier output current flows into the telephone line and a fairly large proportion of the original line voltage is developed between the grid and cathode of the first tube since these circuits are connected to non-conjugate points of the bridge. The output and input terminals of the amplifier are connected to conjugate points of the bridge in such a manner that when the bridge circuit is balanced no feedback is effective at the input.

It has been shown, in the preceding Section, that the negative impedance generated by this form of converter is equal to the network impedance Z_N divided by a conversion factor. To obtain a practical design of the E3 repeater for the faithful conversion of the network impedances, with a minimum of spurious components, it is necessary to balance out, as nearly as possible, all converted circuit elements associated with the output bridge and connections to the line. Accordingly, the two line capacitors are balanced out by a network capacitor. The battery supply resistors, the resistor and capacitor of the plate-battery filter, and the plate-load resistor of the first-amplifier stage are all balanced out in the network by placing suitable values of resistors and capacitors in the cathode circuit of the second-amplifier stage. All of these elements combine to form an equivalent two-terminal network.

An ideal negative impedance device would convert any impedance in the network over a wide frequency band but it is advantageous to limit the negative impedance, in so far as practicable, to the frequency bandwidth required by the particular application. This is accomplished primarily in the network associated with the converter. The conversion bandwidth of the E3 repeater is restricted at the low frequencies by the design of the resistive-capacitive feedback coupling network, between the output bridge and the input grid, and at the high frequencies by the shunt capacitance connected across one resistive arm of the bridge circuit. The circuit is shown in Fig. 18.

The final negative impedance shunted across the line is equal to $-Z_N/0.94$, within ± 2.5 per cent, over the frequency range of 200 to 5,000 cycles. The magnitude and phase of the negative impedances are

controlled by the configuration of the gain-adjusting network, consisting of several inductors, capacitors, and resistors. These components may be arranged in a variety of ways to obtain the gain and frequency shaping characteristics desired for each type of line facility.

The E3 repeater employs a Western Electric 407A twin-triode electron tube of the 9-pin miniature type. The circuit is arranged so that the current for the heater of the tube can be obtained from 24- or 48-volt office battery. The heater power is 2 watts and the plate current is nominally 5 milliamperes.

EQUIPMENT

The objective of the equipment design of the E2 and E3 repeaters was to produce a repeater that would be simpler to manufacture, easier to engineer, install, and maintain, and make more efficient use of the mounting space, particularly on 23-inch bays, than the present E1 repeater. Because of the large demand, savings in manufacturing costs were realized by using a compact aluminum die-cast shell to house the repeater components. To facilitate manufacture, parallel thermoplastic strips were used for the mounting of "pigtail" type of components.⁶ Further savings in shop costs were obtained by coordinating the designs of the E2 and E3 repeaters for maximum interchangeability of parts.

Engineering and installation effort has been reduced considerably by avoiding engineered options and by arranging the equipment so that the maximum portion of the assembly and wiring work is performed in the shop. Testing and maintenance routines are simplified by arranging the repeaters as plug-in units that can be removed from their bay positions and plugged into a portable test set located in a more convenient working space. In this way, the network strapping and any repair work which may be required on a repeater need not be performed while the repeater is in place, or is in a congested area. Maintenance and service interruptions are reduced to a minimum length of time by replacing a defective plug-in unit with a spare for immediate restoration of service, and the faulty repeater repaired at a more convenient time.

REPEATER UNITS

Both E2 and E3 repeater units use the same rectangular die-cast chassis, shown in Fig. 19. The front section of each unit carries the electron tube and test pin jacks which must be accessible for testing and routine maintenance. The gain-adjusting network strapping terminals are arranged, in three rows, along the left side of the repeater, and are accessible only after the unit is removed from its mounting shelf.

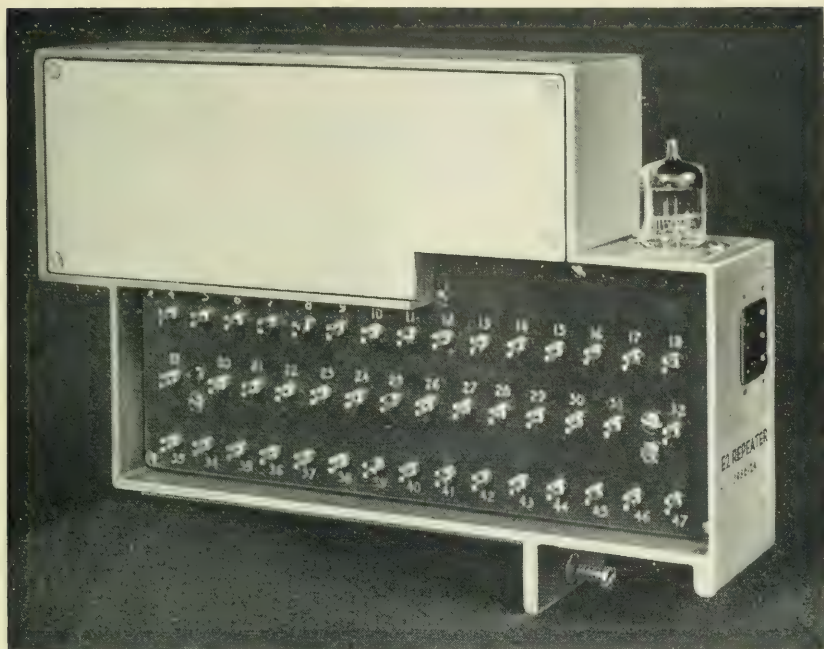


Fig. 19 — Front and side view of E2 repeater.

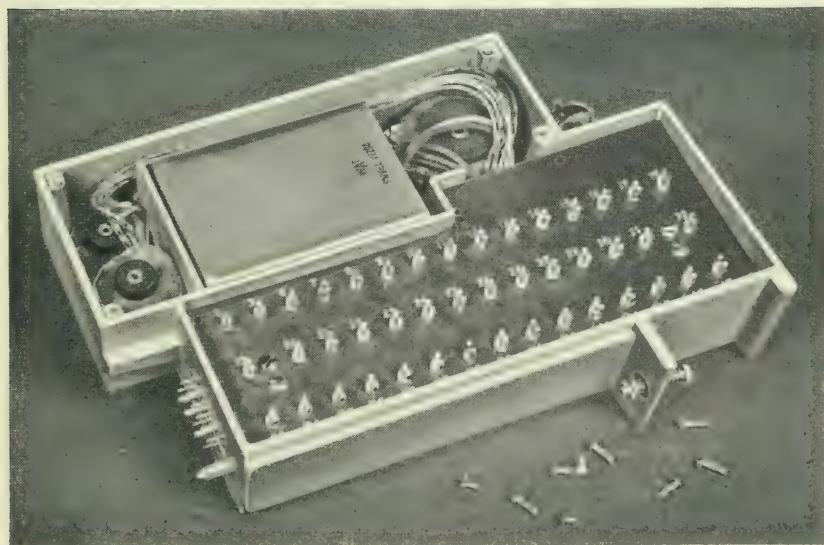


Fig. 20 — Assembly and wiring of E2 repeater.

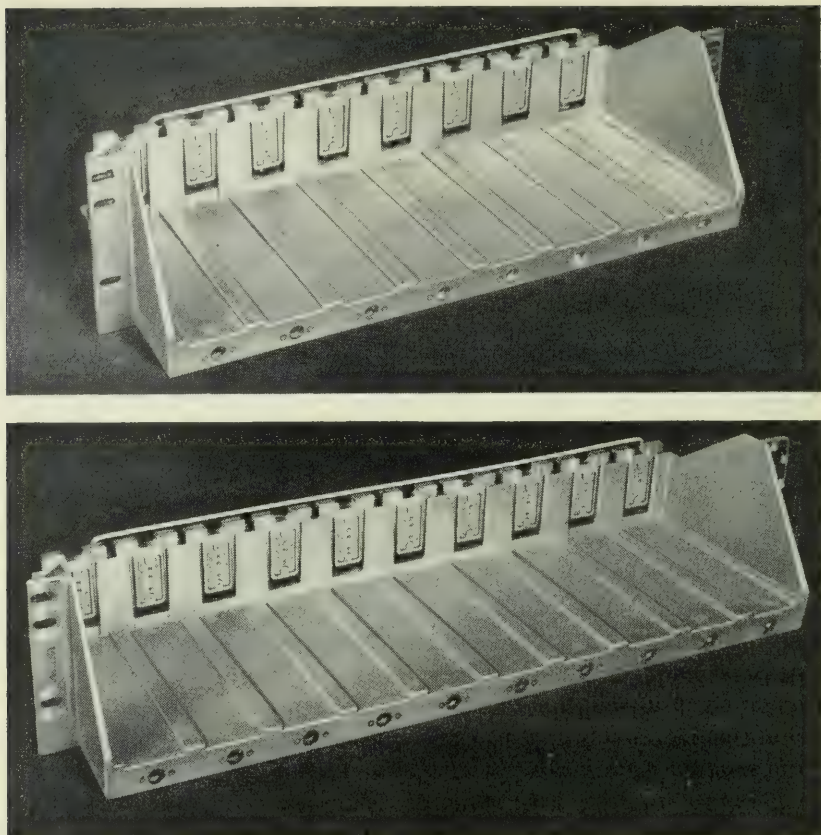


Fig. 21 — Repeater mounting shelves. Upper for 19-inch bays, lower for 23-inch bays.

All external connections, to either type of repeater, are made through a male connector rigidly mounted into the repeater chassis, as shown in Fig. 20. The matching female socket is suspended in the mounting shelf, on a floating assembly, to relieve the strain on contacts and wiring when the repeaters are plugged into a shelf. Both connector units consist of molded rectangular phenolic blocks equipped with 15 gold plated contacts. Proper alignment between the male and female connectors is maintained by using a positioning key and guide pin on the repeater chassis, shown in Fig. 20, and a track on the mounting shelf. After a repeater is seated into its shelf position it is made secure by means of a screwdriver operated quick-acting fastener.

MOUNTING SHELF

An aluminum die-cast shelf, shown in Fig. 21, is used for mounting the repeaters on the relay rack bays. The shelf comes in two widths; one holding 8 repeaters is used for 19-inch bays and the second, containing 10 repeaters, is used on 23-inch bays. Grooves cut into the base of the shelf match a projection on the repeater to act as a track system for positioning the repeater into its connector socket. The shelf connector is mounted in a small die-cast block which, in turn, is fastened to the shelf by shoulder screws to provide a floating assembly. A tapered key, molded into the repeater casting, engages a slot in the connector block to secure the horizontal alignment, and a tapered pin, at the bottom of the repeater casting, raises the connector block for the vertical alignment.

Fifteen mounting shelves can be arranged on standard 11-foot 6-inch relay rack bays. The maximum complement of repeaters will be 120 for 19-inch bays and 150 for 23-inch bays.

POWER DISTRIBUTION PANELS

Fabricated power distribution panels which occupy $1\frac{3}{4}$ -inches of mounting space are used for supplying plate and filament power to the repeater shelves. The primary unit, shown in Fig. 22, is required for furnishing power to the first three repeater shelves and one test-set power outlet. Four sets of plate and filament alarm-type fuses are provided, one set for each shelf and one set for the test power outlet. Four filament-adjusting rheostats are furnished and the panel is equipped with an alarm relay and lamp. Pin jacks are available on the front of the panel to measure the filament voltage.

A supplementary power distribution panel is available, equipped with plate and filament fuses, rheostats and pin jacks sufficient for six additional repeater mounting shelves. One primary and two supplementary power panels are required to fully equip an 11-foot, 6-inch bay with 15 repeater shelves. Both panels are completely wired in the shop to simplify installation. Fig. 23 shows a 19-inch relay bay mounting shelf complete with E2 and E3 repeaters and the two types of power distribution panels.

REPEATER TEST SET

A new test set has been developed, which will simplify the installation and maintenance of these devices. This test set has been designed

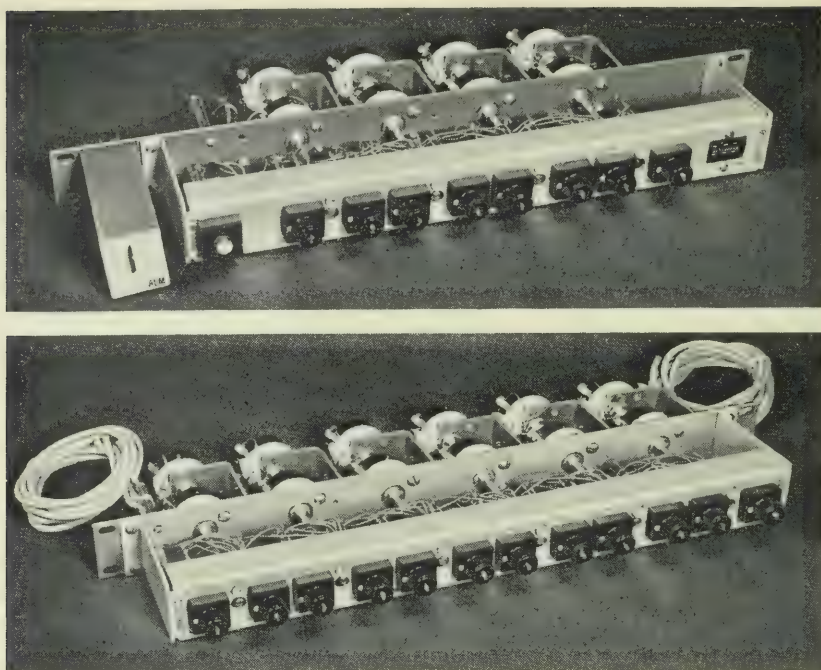


Fig. 22 — Power distribution panels. Upper is primary unit, lower is supplementary unit.

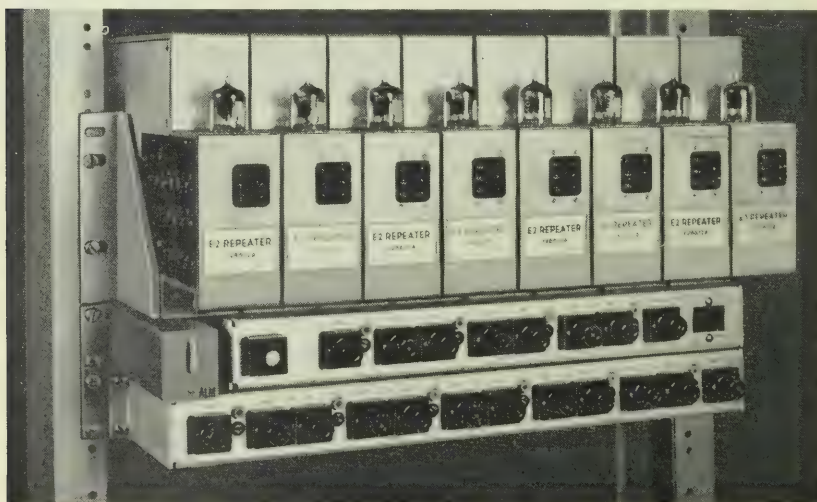


FIG. 23 — Shelf of E2 and E3 repeaters, with power distribution panels, for 19-inch bays.

not only to test the new E2 and E3 repeaters but also the older type of E1 repeater, which has been in use for several years. All test connections which are required to measure individual repeaters or combinations of E1 or E2 and E3 repeaters are made by the operation of a single rotary control function switch. Such a system will avoid the unintentional errors and time consuming operations which attend the setting-up of complicated patch cord connections.

The first five positions of the function switch are arranged to make transmission measurements of an individual repeater or any combination of E1 or E2 and E3 repeaters. Insertion gain and loss measurements are performed with the repeaters working between their normally connected line impedances.

In making transmission and stability measurements, it is sometimes necessary to set up trial connections of the repeater networks during the initial test period or for unusual line conditions. If the actual repeater networks were used, it would require a large number of soldered connections to be made for each condition of the tests. Equivalent jack-ended E2 and E3 networks have been included in the test set, so that with small patch cords, a rapid interconnection of the network components can be made.

As gain and stability of the E-type repeaters are both directly dependent upon impedance variations of the lines, it is sometimes necessary to measure line impedance to determine the causes for low return loss and instability. The last four positions of the function switch are arranged to make positive impedance measurements of lines, negative impedance measurements of the repeaters and return loss measurements between any two lines or impedances. A decade resistance standard has been built into the test set to furnish a direct indication of the magnitude of the unknown impedance. The phase angle of the impedance is determined by a return loss measurement and the angular degrees are read from a chart supplied with the equipment.

Although this test set was designed primarily for tests on E type repeaters, it is possible to connect other types of four-terminal networks into the test set and measure impedances, return loss against known impedances, and insertion gains or losses.

THEORY OF TRANSMISSION MEASUREMENTS

One method of measuring the insertion gain of a negative impedance repeater, without affecting the transmission or stability of the circuit, is to introduce the test voltage, through a low impedance source, in series with the input line and measure the resultant current picked off

through a similar low impedance connection in series with the output line. A comparison of the received currents, with and without a repeater, will give an indication of the insertion gain or loss.

The principles involved in making this type of transmission measurement are shown in Fig. 24. The voltage source and the current measuring detector are assumed to have negligible impedances compared to the impedances of the lines Z_1 and Z_2 .

In the reference condition of Fig. 24, the driving voltage E_0 and the current measuring device are in series with the two line impedances Z_1 and Z_2 . The resultant current will be determined by the source voltage E_0 divided by the vector sum of impedances Z_1 and Z_2 :

$$I_0 = \frac{E_0}{Z_1 + Z_2}$$

where $Z_1 + Z_2$ indicates a vector addition of the two impedances.

In the "measure" condition of Fig. 24 the repeater, represented by the network N , is inserted into the circuit between the voltage source and the measuring device, and, although the voltage E_0 is assumed to remain constant, the addition of network N into the circuit will change

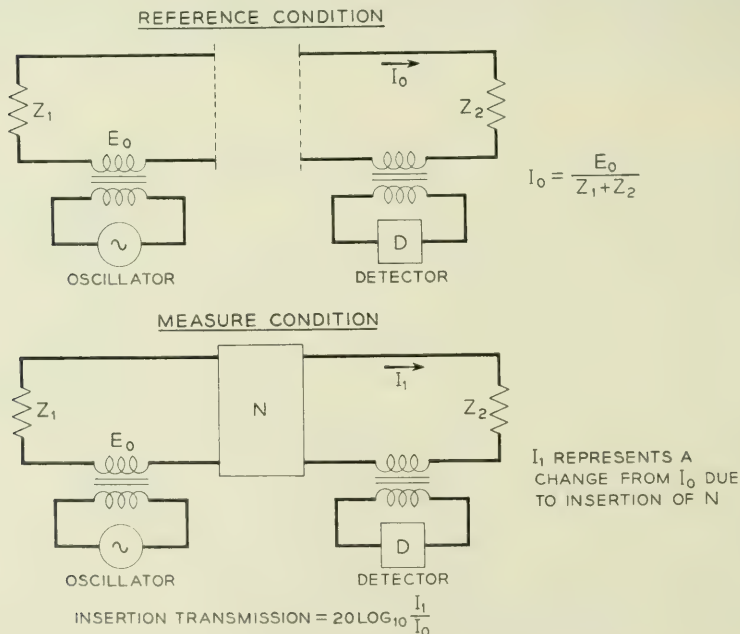


Fig. 24 — Principles of insertion measurements.

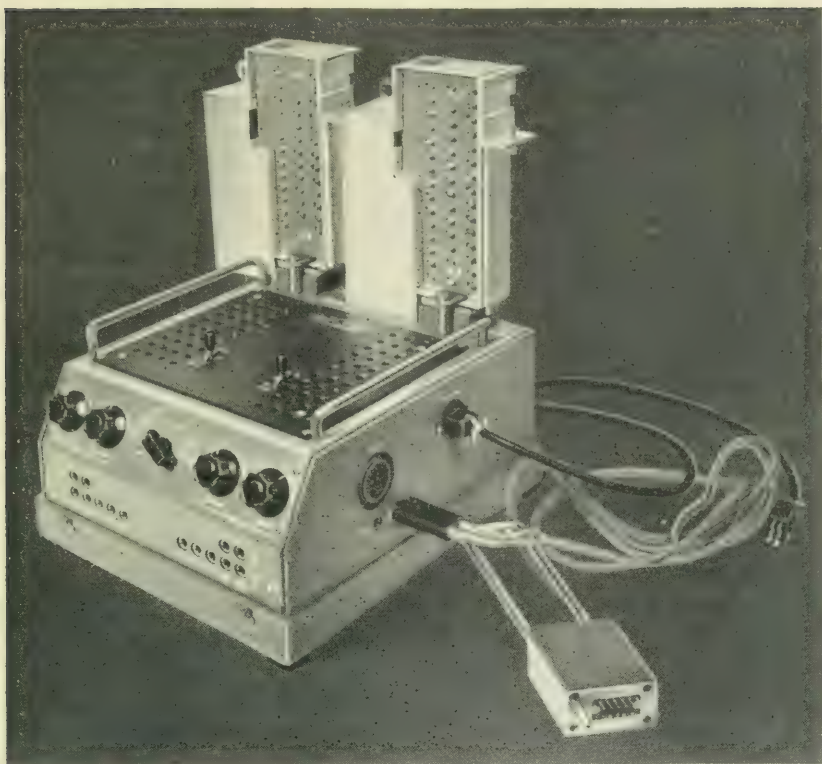


Fig. 25 — Test set showing repeaters and test cords.

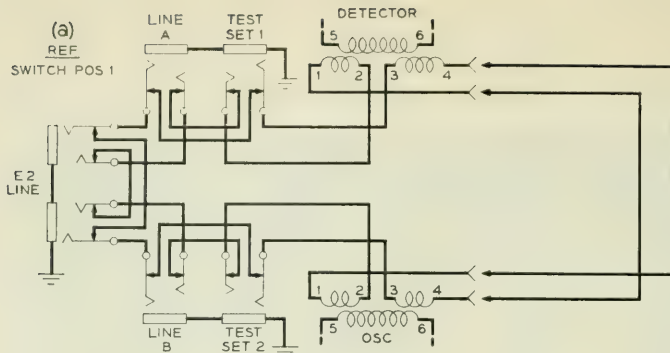
the current to a new value I_1 . The ratio of the currents I_1/I_0 is a direct measure of insertion gain or loss between Z_1 and Z_2 due to inserting the network. The db change in transmission caused by the insertion of N is

$$\text{Insertion transmission in db} = 20 \log_{10} \frac{I_1}{I_0}.$$

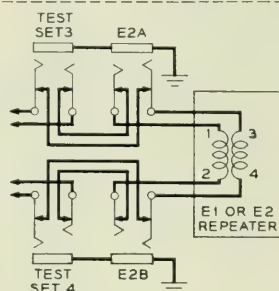
When I_1 is less than I_0 , the addition of N has caused an insertion loss, and when I_1 is greater than I_0 the addition of N has caused an insertion gain. In the test set the current indicating device has a db scale so that the change in transmission can be read directly in db.

TRANSMISSION MEASUREMENTS

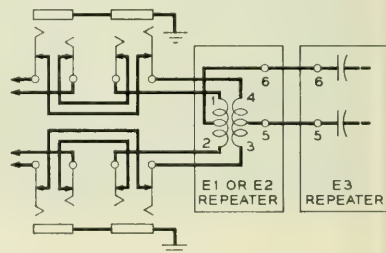
In setting up to make repeater tests, the E2 and E3 units are removed from their shelf positions and plugged into adapters in the test set, as shown in Fig. 25. A test plug, fashioned to simulate the male connector



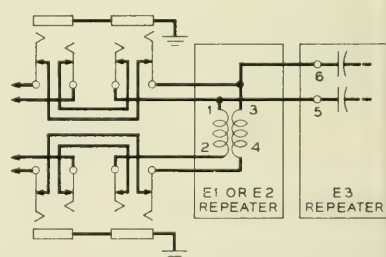
(b)
E1 OR E2
SWITCH POS 2



(c)
E13 OR E23
SWITCH POS 3



(d)
E31 OR E32
SWITCH POS 4



(e)
E3
SWITCH POS 5

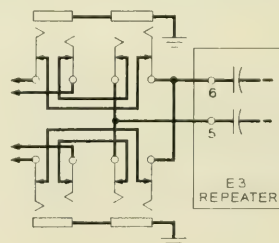


Fig. 26 — Sequence of transmission measurements.

of the repeater, is inserted into the vacant E2 repeater position for picking up the connections to the incoming and outgoing lines. A jack ended cord, connected to the test plug is patched into the test set for applying the line connections to the repeaters under test.

The first position of the rotary function switch arranges the test set for the reference condition. The incoming and outgoing E2 repeater lines provide the terminations, as shown in Fig. 26(a). The low impedance voltage source is obtained by means of an oscillator working through a step-down transformer having an impedance ratio of 600:2 ohms. The low impedance side consists of two equal well balanced windings, one winding being inserted into each side of the line, to form a balanced-to-ground voltage source. The received current is measured by means of a detector working through an identical transformer having one of the 2-ohm windings connected into each side of the line for maintaining the balanced-to-ground circuit.

Switch position 2 connects the E1 or E2 repeater into the test circuit between the sending and receiving impedances as shown in Fig. 26(b). The change in detector reading will indicate the insertion gain or loss of the connected repeater.

The third and fourth switch positions change the arrangements of E1, E2 and E3 repeaters for the insertion measurements of the various repeater combinations, as shown in Figs. 26(c) and (d). Fig. 26(e) indicates the connections for testing the E3 repeater alone on switch position 5.

IMPEDANCE MEASUREMENTS

The second section of the type-E repeater test set has been arranged to measure the impedance of two-terminal networks and the return loss between any two impedances or between an unknown impedance and a specified network. The impedance measuring circuit consists of a hybrid coil arranged in the form of a balanced bridge for measuring return loss. The driving voltage and detecting device are connected across conjugate arms of the bridge and the unknown impedance and a resistance standard are applied across the opposite conjugate arms of the bridge circuit.

Simple transmission measurements made across the hybrid coil, between the oscillator and detector sides of the bridge, are the only determinations required to find the return loss between any two impedances or the magnitude and phase angle of an unknown impedance. Such a device has several advantages over other types of impedance bridges in that no critical balances are required, no calibrations are

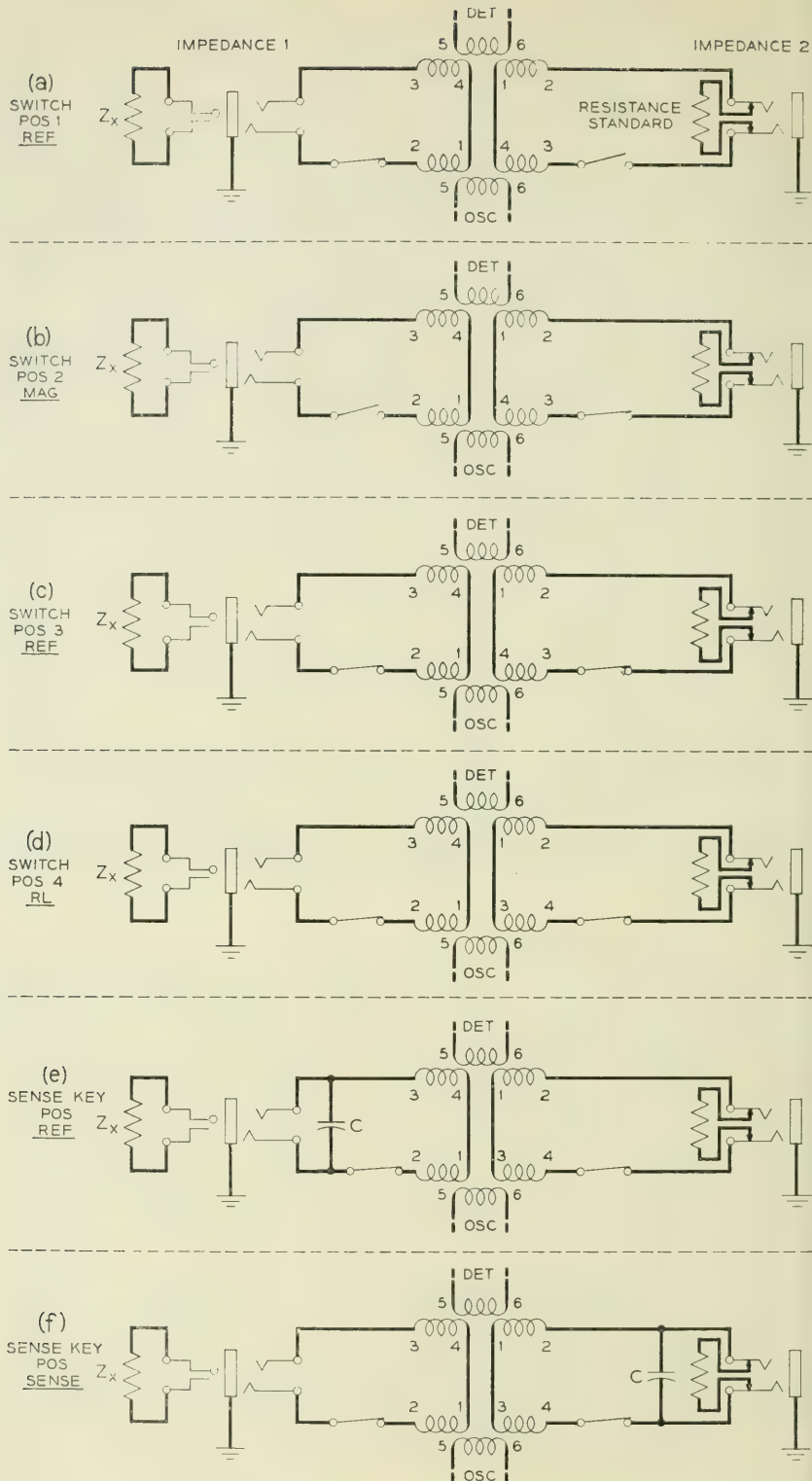


Fig. 27 — Sequence of impedance measurements.

necessary, the critical elements are stable passive networks and changes in the sensitivity of the measuring device, which includes variations in oscillator output and sensitivity of the detector circuit, will not affect the accuracy of measurement.

An impedance reference is established, as shown in Fig. 27(a), when the transmission through the hybrid network is measured with connection to the resistance standard. A second transmission measurement is made through the hybrid coil network with the standard impedance connected instead of the unknown. After replacing the unknown with the resistance standard, the transmission through the hybrid coil is adjusted by varying the resistance standard to obtain the same value of loss, as measured with the unknown impedance. When the two transmissions are equal, the resistance in ohms, as read on the resistance standard, will be the same as the magnitude of the unknown impedance. With the magnitude of the resistance standard and unknown impedance the same, the phase angle of the unknown is readily determined by comparing the transmission through the hybrid coil for two polings of one of the hybrid windings. The first poling, Fig. 27(c), which is used as a reference, provides a measure of the transmission through the hybrid network when the current in the unknown branch and the current in the resistance branch are added vectorially in the detector winding. The reverse poling of one of the coil windings, Fig. 27(d), provides a measure of the transmission through the hybrid when the currents in the resistance and unknown branches are subtracted vectorially in the detector circuit. In addition to being a method of measuring the phase angle of an impedance, this comparison of two measurements with different poling is a return loss measurement of an unknown against a known impedance of equal magnitude. A curve of phase versus return loss may be used for convenient interpretation of the return loss measurement in terms of phase angle. After the phase angle has been ascertained it is necessary to determine the sense of the angle, that is, whether the unknown impedance contains positive or negative reactance. A reference condition is established, Fig. 27(e), by shunting a small value of capacitance first across the unknown impedance and then across the resistance standard and noting the change in transmission through the hybrid network. An increase in the transmission, in going from the reference to measure conditions, indicates a positive reactance and a decrease in transmission indicates a negative phase angle. The reference condition is established to minimize indicational errors at small phase angles of the unknown impedance.

Two additional pieces of information are revealed in the return loss

measurement. When the return loss is positive, the unknown impedance contains positive resistance and the phase angle must lie between 0 and ± 90 degrees. For negative return losses the unknown impedance contains negative resistance and the phase angle must lie between ± 90 degrees and 180 degrees.

ACKNOWLEDGMENTS

The design of the E3 repeater was based upon theoretical work by S. T. Meyers; the design of the test set was based upon theoretical work by H. Kahl and S. T. Meyers. In regard to the theory of negative impedances the contributions of F. B. Llewellyn are gratefully acknowledged.

REFERENCES

1. J. L. Merrill, Jr., A Negative Impedance Repeater, A. I. E. E. Trans., **69**, Part 2, pp. 1461-1466; 1950.
2. J. J. Pilliod, Fundamental Plans for Toll Telephone Plant, A. I. E. E. Trans. **71**, pp. 248-256.
3. O. B. Blackwell, Time Factor in Telephone Transmission, A. I. E. E. Trans., **51**, pp. 141-7.
4. G. Crisson, Negative Impedance and The Twin 21-Type Repeater, B. S. T. J., **31**, pp. 485-513, July, 1931.
5. J. L. Merrill, Jr., Theory of The Negative Impedance Converter, B. S. T. J., **51**, pp. 88-109, Jan., 1951.
6. W. E. Kahl and L. Pedersen, Some Design Features of the N-1 Carrier Telephone System, B. S. T. J., **51**, pp. 418-446, April, 1951.

Stress Systems in the Solderless Wrapped Connection and Their Permanence

By W. P. MASON and O. L. ANDERSON

(Manuscript received December 16, 1953)

The solderless wrapped connection is initially held together by the hoop stress in the wire which enters the connection as a result of the tension put on the wire by the wrapping tool. Measurements made out to a time of 1.5 years at room temperature show that the tension has decreased to 70 per cent of the one day value (8000 lbs per square inch) in this period. Two methods of extrapolation are discussed, both of which indicate that at least half of the initial one day value will remain at the end of forty years at room temperature.

Another set of stresses enters the connection as a function of time, namely the diffusion forces produced by diffusion of the tin plating into the brass terminal and copper wire. A number of experiments are discussed which show that the activation energy of diffusion is materially reduced by the shearing stresses in the connection. Measurements at two temperatures, which allow extrapolation to room temperature, indicate that at the end of two years the force required to strip the wire from the terminal has increased by 5 per cent over the initial value and that at the end of forty years the increase will be 20 per cent. Support for these conclusions is furnished by tests on actual connections that have been in the field for one year and ten months, which show an increase of 5 per cent in the stripping force even though the relaxed hoop stress is only 68 per cent of the initial value. The increase, which is due to diffusion forces, can be made higher by using zinc, cadmium or aluminum plating, and the fusion occurs in a shorter time.

INTRODUCTION

As discussed in a series of papers,¹ the solderless wrapped connection is an efficient and inexpensive method of connecting a wire to a terminal. All the tests made so far indicate that it is mechanically sound and sufficiently free from the effects of corrosion to have a trouble free life of at least forty years. Photoelastic and stress studies show that the

¹ The Solderless Wrapped Connection, B. S. T. J., **32**, May 1953.

connection is initially held together by the hoop stress in the wire which enters the joint as a result of the tension put on the wire by the wrapping tool. As time goes on, another system of stresses are generated, namely, the diffusion stresses caused by the diffusion of one part of the connection into the other which eventually eliminate the surface between the wire and the terminal and in effect join the two together.

It is the principal purpose of this paper to describe a number of experiments which show how these stresses develop, how large they are and how they can be increased by substitution of a different type of plating between the terminal and the wire. A comparison of the strength of a joint formed by a tin plated copper wire and a bare copper wire both wound on nickel silver or brass terminals shows that the diffusion forces develop more quickly when the wire is tin plated than when it is bare. Measurements made at different temperatures show that the activation energy of diffusion is decreased in proportion to the hoop stress in the wire indicating that the shearing stress at the contact surfaces aids diffusion. This activation energy is considerably less for tin than for copper. The diffusion joint has a strength per unit area equal to the limiting shearing stress of the tin plating. This is in the order of 3,000 pounds per square inch for tin but is considerably higher for other types of plating such as zinc, aluminum or cadmium. Measurements of the stripping force of connections made with bare and tinned copper wire on zinc, aluminum or cadmium plated terminals show that the stripping force increases by a factor of two as a function of time, and the time required for the diffusion forces to operate is considerably less with these types of plating.

The combination of relaxation and diffusion stresses that are discussed later show that as the mechanical strength due to the hoop stress decreases, the strength due to diffusion increases and at the end of forty years the standard tin plated wire on nickel silver or brass terminals will be at least 20 per cent stronger than it is initially. The extensive corrosion tests described in Footnote 1, taken together with the mechanical strength tests described here, show that the standard connection should not fail in the forty year period under consideration.

RELAXATION OF HOOP STRESS AS A FUNCTION OF TIME

The rate of relaxation of the hoop stress in the wrapping wire is an important quantity for the stability of the connections. This has been studied by wrapping 24 gauge wire with a constant tension of three pounds around a spring steel terminal 0.0124 inches thick and 0.062 inches wide. As shown by Figure 11 of Mallina's paper,² this causes the

terminal to twist through an angle of about 25° when 100 turns are wrapped around the terminal. This twist is the result of a torque which is caused by the fact that the wire has a hoop stress and the wire does not come back on itself but advances by the thickness of the wire for each turn. As shown previously,² the total torque is equal to

$$\text{Torque} = (\text{W.F.}) \left(\frac{2ab}{a+b} \right) L \quad (1)$$

where (W.F.) is the wrapping force, $2a$ and $2b$ the thickness and width of the terminal and L the total length of the wrapped section. Hence, by calibrating the spring constant, the average hoop stress can be evaluated. In this manner, Mallina² has found that after transient creep (defined

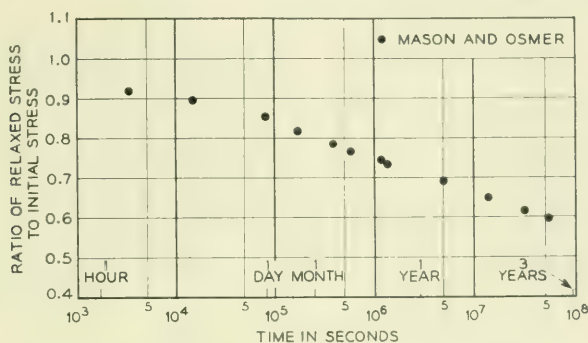


Fig. 1 — Relaxation of stress in copper solderless wrapped connection at room temperature.

in this paper as the loss in stress during one day) has occurred, the stress level for the standard 24 gauge connection is about 8,000 pounds per square inch.

This same technique can be used to measure the loss in hoop stress as a function of time for all one has to do is to put a pointer on the spring and observe the rate the angle decreases. It will be noted that this system reproduces all of the stress systems and strain hardening that occur in the wrapped solderless connection, and hence a measurement of the relaxed angle is the most representative measurement of the relaxation of the solderless wrapped connection. Fig. 1 shows an average of the results of four such springs from the time of their initial formation out to a period of one year and one month. The rate of relaxation is quite rapid during the first day (transient creep) but at the end of one

² R. F. Mallina, Solderless Wrapped Connections, Part I — Structure and Tools, B. S. T. J., pp. 525-555, **32**, May 1953.

year and one month the average hoop stress is 5,900 pounds per square inch and it is decreasing at the rate of 240 pounds per square inch between the first and second year. In the previous paper,³ since the transient creep occurring during the first day is somewhat variable, the angle of twist reached 24 hours after the connection had been made, was taken as the initial 100 per cent value and the relaxation has been plotted with respect to this. Fig. 2 shows this plot with data from Mason and Osmer and a set of eight relaxing springs measured by Mallina and McKettrick all plotted on the curve. At the end of 1.5 years, the observed stress is 70 per cent of 8,000 pounds per square inch or 5,600 pounds per square inch on the average.

The measurements out to 1.5 years by this method appear to be the longest measurements of stress relaxation in annealed copper wire in the

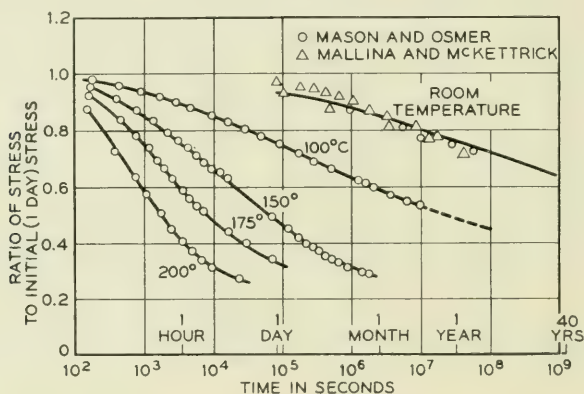


Fig. 2 — Relaxation of stress in tinned copper wire as a function of time and temperature.

literature. Since the wrapped solderless connection is expected to last at least forty years, it is desirable to be able to extrapolate the stress value out to forty years time. Several methods are possible. According to theoretical treatment,⁴ stress relaxation proceeds as follows:

$$\sigma_0 - \sigma = \frac{kT}{\beta} \log_e (1 + t/t_0) \quad (2)$$

³ W. P. Mason and T. F. Osmer, Solderless Wrapped Connection, Part II — Necessary Conditions for Obtaining a Permanent Connection, B. S. T. J., **32**, pp. 557-590, May 1953.

⁴ See *Progress in Metal Physics*, Volume IV, Pergamon Press, Limited, 1953, Chapter V, Theory of Dislocations by A. H. Cottrell, pp. 233 and 251 to 260; Kuhlmann, D. and Z. Physik, **24**, p. 43, 1952; and Proc. Phys. Soc., **64**, p. 64, 1951; and A. H. Cottrell and V. J. Ayetekin, Inst. Metals, **77**, p. 389, 1952. The last reference gives data for stress relaxation in zinc which shows that it obeys equations (2), (3) and (4).

where σ_0 is the initial stress, σ the final stress, k Boltzmann's constant, T the absolute temperature, t_0 a time determined from the equation

$$\sigma_0 = -\frac{kT}{\beta} \log \left(\frac{\beta A t_0}{kT} \right) \quad (3)$$

and A and β are constants in the fundamental equation

$$\frac{d\sigma}{dt} = -Ae^{-(H-\beta\sigma)/kT} \quad (4)$$

where H is the activation energy in the absence of any stress.

The result of (2) is that for times large compared to t_0 , the stress difference $\sigma_0 - \sigma$, plotted on logarithmic paper will be a straight line relation. Hence, in Fig. 2, for times in the order of forty years, the indicated stress level is 60 per cent of 8,000 pounds per square inch or 4,800 pounds per square inch.

Another method of extrapolation which essentially makes use of (4) is to measure the rates of relaxation at different temperatures and determine an activation energy for each stress level by comparing the logarithms of the times for the same stress levels as a function of temperature. The constant A in (4) is usually considered to have a temperature factor T in it and is usually written as $A'T$. With this relation, (4) can be written

$$-\frac{d\sigma}{A'Te^{-(H-\beta\sigma)/kT}} = \frac{-e^{-(H-\beta\sigma)/kT}}{A'T} d\sigma = dt \quad (5)$$

Integrating this equation between the limits σ_0 and σ , we have

$$t = \frac{k}{\beta A'} e^{H/kT} [e^{-(\beta\sigma/kT)} - e^{-(\beta\sigma_0/kT)}] = \frac{k}{\beta A'} e^{(H-\beta\sigma)/kT} [1 - e^{-\beta(\sigma_0-\sigma)/kT}] \quad (6)$$

Hence, if the quantity

$$e^{-\beta(\sigma_0-\sigma)/kT}$$

is small compared to unity, we have

$$(H - \beta\sigma) = k \log_e (t_1/t_2) \bigg/ \left[\frac{1}{T_1} - \frac{1}{T_2} \right] \quad (7)$$

As shown by Fig. 3, the value of β is about 6 calories per pound per square inch and hence if $\sigma_0 - \sigma$ is at least 800 pounds square inch (i.e., for all values of relaxation of 0.9 or less) this factor will be less than 1 per cent for all temperatures used and hence we can use (7) to evaluate values of the activation energy $H - \beta\sigma$.

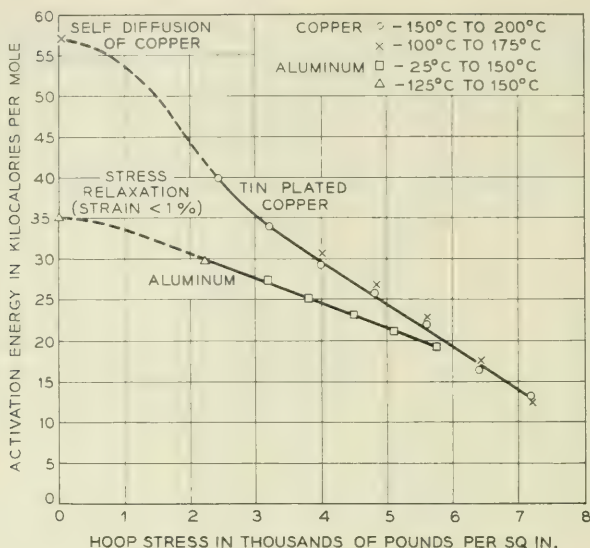


Fig. 3 — Activation energy for stress relaxation as a function of hoop stress.

Measurements of the rate of stress relaxation for tinned copper wire at 200°C, 175°C, 150°C, and 100°C are shown by Fig. 2, and the activation energies plotted against average hoop stress are shown by Fig. 3. Values are given using 150°C to 200°C as the temperature range and 100° to 175° also. Both ranges give the same activation energies within the experimental errors and show that down to about 0.4 relaxation the activation energies satisfy an equation of the type

$$H' = H - \beta\sigma \quad (8)$$

The curvature exhibited by the relaxation versus $\log t$ shown for all temperatures indicates that the activation energy must increase faster than (8) for low values of relaxation and the dotted line shows a hypothetical curve ending up at the self diffusion activation energy for copper, 57 kilocalories per mole.

To apply this method in general, one has to take account of any transformation such as recrystallization in the temperature range of measurement. For example, Fig. 4, shows similar curves for aluminum wire. Recrystallization in aluminum is known to occur at temperatures above 150°C and this change is shown in the relaxation measurements by the lower values of relaxation that occur for long times. If one takes values of time above and below the recrystallization temperature, the activa-

tion energy will appear higher for this range than for a temperature range below the recrystallization temperature. The agreement of the activation energy for copper for the two temperature ranges shown by Fig. 3, shows that no transformation occurs from 25°C to 200°C and hence we can extrapolate the relaxation to room temperature taken as 25°C, with the result shown by the solid line labelled 25°C. This agrees well with the measured values and indicates a stress at the end of forty years equal to 5,200 pounds per square in.

DIFFUSION STRESSES IN SOLDERLESS WRAPPED CONNECTIONS

In addition to the hoop stresses, another set of stresses develops as a function of time, namely, the diffusion stresses caused by the diffusion of one part of the connection into the other. The first experiment that showed the presence of these stresses was the stripping force tests of Fig. 23 of the previous paper referred to in Footnote 3. These measurements were carried out on connections which had been held at 175°C for lengths of time up to ten days and it was found that the stripping forces did not decrease with time. A more careful set with twenty connections for each point have recently been run with the results shown by Fig. 5, solid line labelled 175°C. From this curve, it is seen that the stripping force decreases to 88 per cent of its initial value of 15.5 pounds average for six turns of 24 gauge tinned wire and then increases to 120 per cent of the initial value at the end of ten days. Similar increases are shown at 100°C over a longer period of time and recent tests of solderless wrapped connections that have been in the field for one year and ten months show that the stripping force is about 5 per cent higher on the average than it was when the connection was formed.

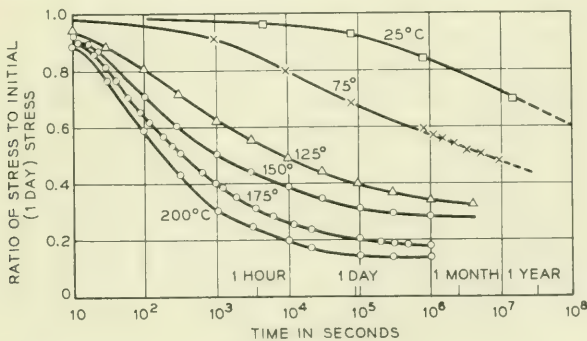


Fig. 4 — Relaxation curves for aluminum wire in solderless wrapped connections.

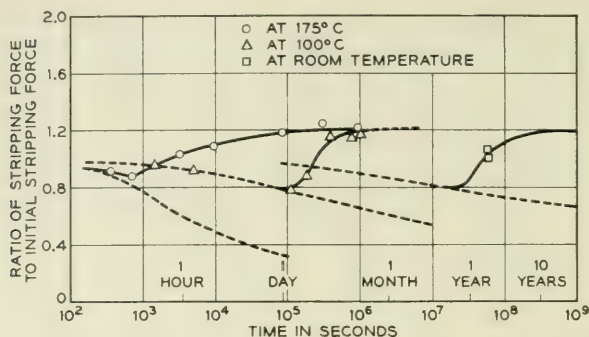


Fig. 5 — Ratio of measured stripping force to initial stripping force as a function of time and temperature for tinned copper wire on nickel silver terminals.

The dashed lines show the corresponding hoop stresses as a fraction of the initial hoop stress and hence it is evident that as the hoop stresses go down the stripping forces first decrease and then increase to higher values than were effective originally. We shall presently show that the difference between the measured stripping force and the proportionally relaxed hoop stress is due to a stress caused by diffusion of the tin in the plating into the wire and terminal of the solderless wrapped connection.

One experiment which shows that the initial stripping force is caused by the hoop stress in the wire is the experiment shown by Fig. 6. Here a terminal is made which has a slightly tapered tin plated pin in the middle and is cut back for some distance beyond the pin. The terminal is wound with five turns of No. 14 gauge (0.065 inch diameter) tinned copper wire. The initial stripping force to pull off the winding was determined and it was found to take 72 pounds force on the average to strip the wire off the terminal. A similar set of measurements was made on the force required to pull the pin out of the terminal and this averaged about 50 pounds or 70 per cent of the stripping force for the wire. Since the pin and terminal had the same coefficient of friction as the wire and terminal, and since the pin is required to support all the compressional stress in the terminal due to the hoop stress in the wire, it is evident that at least 70 per cent of the force required to strip the wire off the terminal is plain frictional force between the wire and the terminal. It is thought that the remainder of the force is due to the gouges cut in the terminal by the winding process. When the wire is stripped off the terminal, these cuts gouge out parts of the wire and hence require a higher force. This effect is equivalent to friction for a rough surface, which is higher than that for a smooth surface.

Next the terminal was held at 175°C for various times as shown by Fig. 6, where the stripping forces and the forces to pull the pins are plotted. It is seen that the force required to pull the pin decreases with time while the force required to strip the terminal increases with time. The pin force duplicates the stress relaxation curve initially but departs from this curve more and more as time progresses. This gradual departure appears to be due to some diffusion in the tin which makes partial contact with the terminal. The amount of diffusion between pin and terminal is less than that between wire and terminal for two reasons: (a) the contact area between pin and terminal is not as intimate as the contact area between wire and terminal due to excessive plastic flow in the latter but not the former case, and (b) the contact area between pin and terminal is much greater than between wire and terminal. Nevertheless, the force required to pull the pin is substantially the same as the frictional force holding the wire on the terminal due to the hoop stress in the wire. Hence, we can conclude from this experiment that the initial stripping force is due to the frictional force resulting from the hoop stress plus shearing forces required to gouge the wire.

Several experiments have been undertaken to measure the effect of diffusion separate from friction. The most successful of these was the arrangement shown by Fig. 7. Here a wire was pressed against a double-toothed sharp edged block, and a constant pressure was maintained for various times at various temperatures. An attempt was made to produce an indentation of the same magnitude as that developed in the wrapped solderless connection, although, of course, the area of contact increased slightly with time.

It was found that for aluminum or copper on nickel silver at room

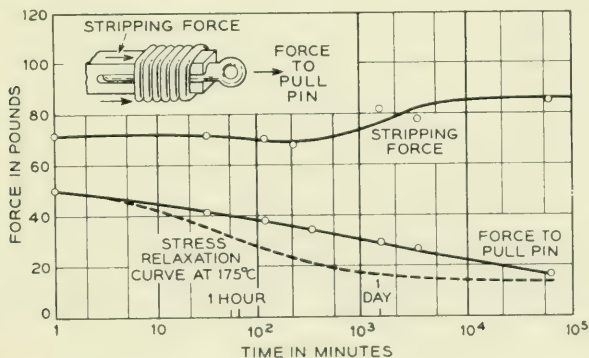


Fig. 6 — Experiment showing difference between stripping force and frictional force due to hoop stress — temperature 175°C .

temperature, no adhesion occurred even when the time of loading was very long. It was also found that if the load was removed, even momentarily, before the sample was placed in the oven, no adhesion occurred. For a given temperature there is an induction period before the wire adheres to the block, and this induction period increased, in general, as the temperature decreased. This induction period was found to be a time of nucleation. This is shown by the fact that the period increases as the square of the contact dimensions. Since diffusion is a function of $x/\sqrt{Dt}$ where x is the distance, D the diffusion constant and t the time, this observation shows that nucleation starts at a given point and proceeds for a certain fraction of the contacting surface before fusion strengths are observed. After the induction period, the fusion force — the force required to pull wire and block apart — increased at a rapid rate for the cases of copper, tinned copper, and zinc wire on a nickel silver block.

In order to account for the effect of fusion due to the increase of contact area, the ratio of the fusion force to the contact area was determined, yielding a shearing stress. The area of contact could be easily measured with a microscope since a bright surface was produced by the shearing process. The shearing stress, Fig. 8, for the case of tinned copper wire on nickel silver is shown to approach a limit of about 3,000 psi which is approximately the limiting shearing stress for tin. Hence, it appears that tin diffuses into the copper wire and the nickel silver base. Further-

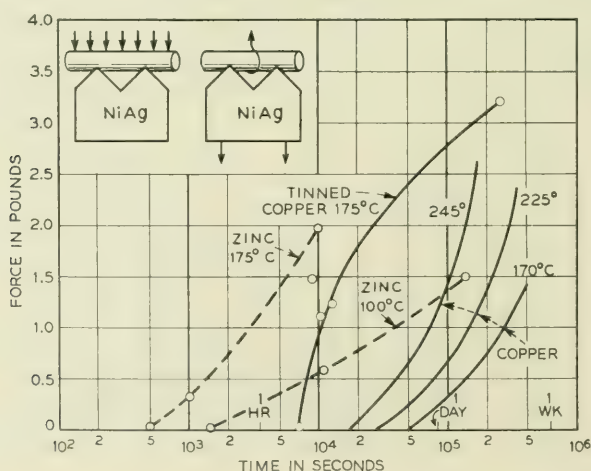


Fig. 7 — Diffusion forces for tinned and bare copper and zinc wire on nickel silver base.

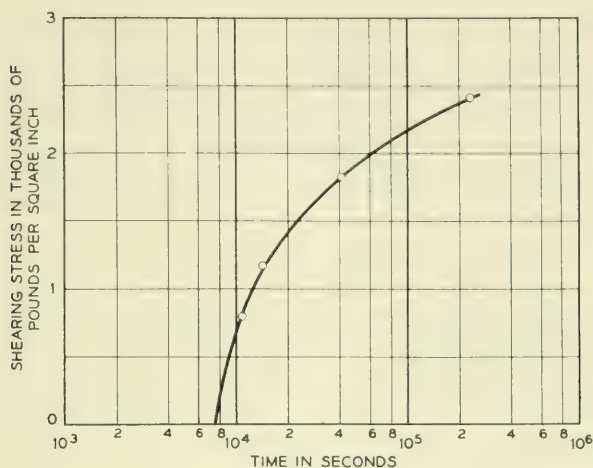


Fig. 8 — Diffusion shearing strength for tinned copper wire on nickel silver at 175°C as a function of time.

more, it appears that the fused bond will just withstand a shearing stress equal to the limiting shearing stress of tin. Metallurgical evidence for this diffusion was presented in Fig. 19 of the paper referred to in Footnote 4. In this figure, a tin plate layer between the copper and nickel silver was diffused in the two metals in a time of about 400 hours at 180°C . The new evidence indicates that this layer of tin has the shear strength of bulk tin.

The previously described data constitutes measurements of fusion forces separate from the complications of friction in the solderless wrapped connection. However, the same measurements of the effect of fusion are obtained if one subtracts from the stripping force the product of the relaxed force times the initial stripping force (about 15.5 pounds) of the actual wrapped connections. As shown by Fig. 6, the fusion force so computed divided by the area of contact of a six turn connection (about 0.0045 sq in) yields the shearing stress, previously defined. In Fig. 9, the shearing stress is given for a tinned copper wire on a nickel silver terminal as a function of time for 175°C and 100°C . By way of comparison, the constant stress diffusion process of Fig. 8, is shown plotted by the dashed line on this plot. It is evident that there is a larger induction period in the case of the 14 gauge (0.065 inch wire) than in the case of the smaller 24 gauge (0.020 inch) wire and the ratio of the times is proportional to the square of the wire diameter ratio. This is an indication that we are dealing with a nucleation process.

The previous results show that the effect of temperature above room temperature on the solderless wrapped connection is to promote fusion which brings an additional set of forces into the picture with regard to the mechanical stability of the connection. In order to determine the effect of fusion at room temperature, it is necessary to evaluate the activation energy of diffusion. This is done by calculating the value H in the simple rate equation

$$\tau = \tau_0 \exp (H/kT)$$

so that a given change of temperature corresponds to a given change of time on the fusion force curve. If we compare the 175°C curve with the 100°C curve of Fig. 9, an activation energy curve as a function of residual hoop stress is obtained as shown by Fig. 10. The value of stress is determined by the average time values used in the calculation of the activation energy. It is shown that the activation energy for diffusion of tinned copper on nickel silver in the presence of stress is in the order of 20 to 24 kilocalories per mole while that with no stress is in the order of 41 kilocalories per mole. It appears that higher stresses lower the activation energy of diffusion just as they lower the activation energy for creep and stress relaxation. Very high stresses can reduce the activation energy to a value approaching zero which is the case of cold welding. The fusion occurring in the case of the solderless wrapped connection is different from cold welding in that it takes a finite time to develop any fusion forces.

From the activation energy values of Fig. 10, it is possible to predict the rate at which the wrapped wire connection increases its strength at room temperature. Such a calculation applied to the case of tinned copper wire on a nickel silver terminal is given in Fig. 5 where it is shown

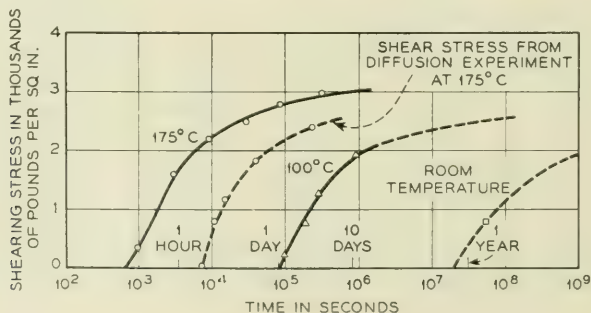


Fig. 9 — Shearing strength of copper-tin-nickel silver connections as a function of time and temperature.

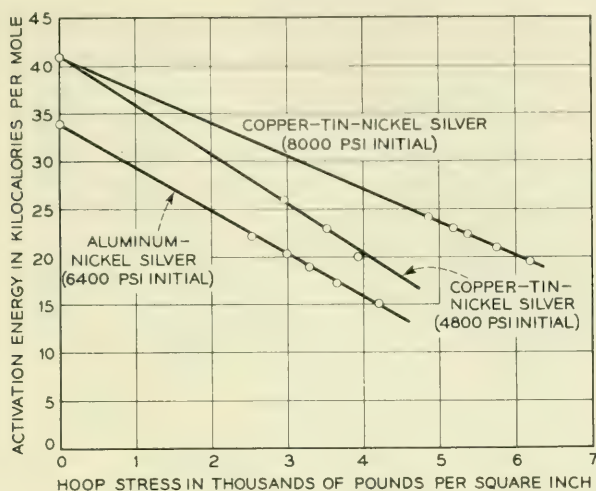


Fig. 10 — Activation energy for diffusion as a function of hoop stress.

that the strength stops decreasing at the end of six months and eventually increases to 20 per cent over the initial value at the end of forty years. Measurements on connections made one year and ten months ago, taken by V. F. Bohman, confirm this calculation.

In all probability, the stress causing the diffusion is the shearing stress applied to the tin layer by the hoop stress of the wire acting on the fixed terminal. For a given initial winding stress, this shear stress will decrease in proportion to the relaxed hoop stress as confirmed by the data of Fig. 10. If we start with a lower winding stress, a smaller indentation is made in the wire and the area of contact is smaller about in proportion to the ratio of the hoop stresses. Hence, although the compressional force on the corners is smaller in proportion to the hoop stress, it is supported by a smaller area and hence the stress on the supporting area is approximately independent of the initial value of the hoop stress. This compressional stress produces shear stresses on the layer of tin since the material of the terminal and the wire have to slide with respect to each other in producing the indentation. Hence, we should expect that the diffusion forces would start at a time independent of the value of the initial hoop stress.

This supposition is confirmed by the data of Fig. 11, obtained by winding tin plated copper wire on a nickel silver terminal with a hoop force half of the usual winding force of 1,300 grams for a 24 gauge wire. The initial strip off force was 60 per cent of that for a 1,300-gram winding

force confirming the data of Fig. 16 of Mallina's paper, referred to in Footnote 2, which shows that the hoop stress decreases less rapidly than the applied tension. Fig. 11 shows the strip off force as a function of time and temperature. If we subtract the relaxed frictional force from the stripping force and divide by the area of 0.0027 square inches (60 per cent of the area of the usual connection) the shear strength curves in pounds per square inch are shown by the lower solid curves. These start sooner than the values on Fig. 9 but require a somewhat longer time to reach their final value. The activation energy obtained from these curves is plotted on Fig. 10 (4800 psi) and shows that the activation energy for the lowest stress is about the same as that for the higher

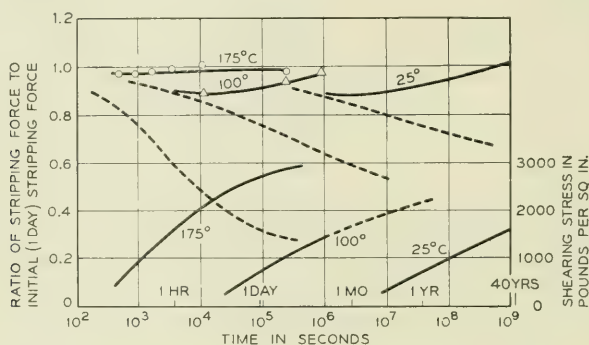


Fig. 11 — Stripping force and shear strength of copper-tin-nickel silver connections as a function of time and temperature. Initial stress is 4,800 pounds per square inch.

stress curve, showing that the shear stress is independent of the wrapping tension. It takes longer to complete the process and as a result, the stripping force at the end of forty years is equal to the initial value rather than 20 per cent higher as in the case for Fig. 5. Corrosion tests for these lower stripping force values⁵ show, however, that there is no indication of a loss of electrical contact.

The method for analyzing fusion forces by subtracting relaxation data from stripping force data is verified by measurements on other metal connections. Fig. 12 shows the ratio of stripping force to initial stripping force for aluminum on nickel silver terminals as a function of time and temperature. Subtracting the relaxed force (the original stripping force multiplied by the relaxed stress ratio of Fig. 4) from the stripping force and dividing this difference by the contact area, the shearing

⁵ R. H. Van Horn, Solderless Wrapped Connections, Part III — Evaluation and Performance Tests, B. S. T. J., **32**, May 1953. See Table II.

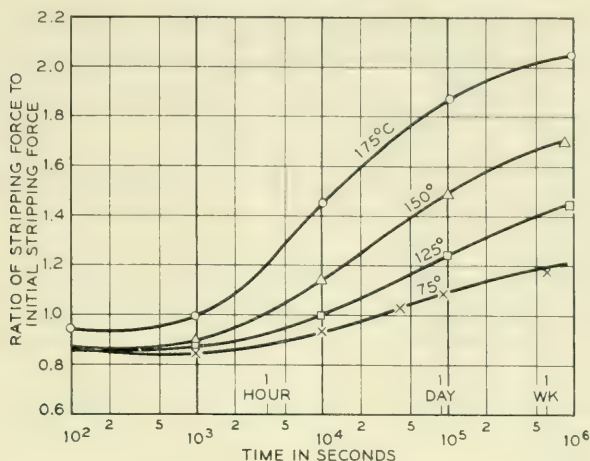


Fig. 12 — Stripping force curves for aluminum on nickel silver as a function of time for four temperatures.

strength is obtained and is shown in Fig. 13. The shearing strength thus obtained approaches 6000 psi for long times at high temperatures. This value agrees approximately with the accepted value for bulk aluminum. Calculating the activation energy of diffusion as previously outlined, one obtains the curve for aluminum in Fig. 10.

The advantage of a diffusion layer such as tin or aluminum as an element in the solderless wrapped connection is shown by the data of Fig. 14 which shows the rate of increase of the stripping force of bare

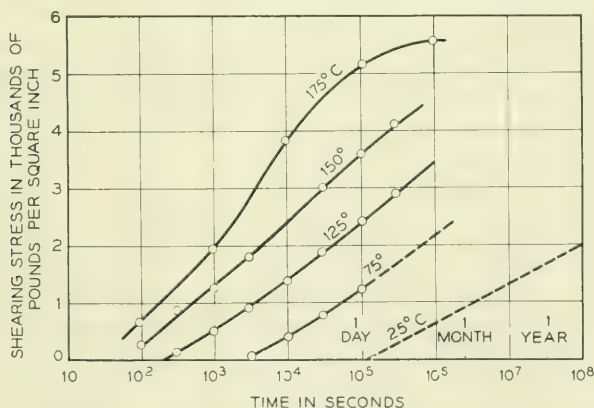


Fig. 13 — Shearing strength of aluminum on nickel silver connections as a function of time.

copper wire on nickel silver terminals at 150°C and 100°C. The increase in strength does not occur as fast as that for tin plated copper, and hence the standard connection with a tin plated copper wire is better than one formed from bare copper wire.

PROPERTIES OF SOLDERLESS WRAPPED CONNECTIONS FOR OTHER TYPES OF PLATING

A study of the factors governing the diffusion strength of the solderless wrapped connection suggests methods for increasing the rate of diffusion and the strength of the diffusing layer. These methods result

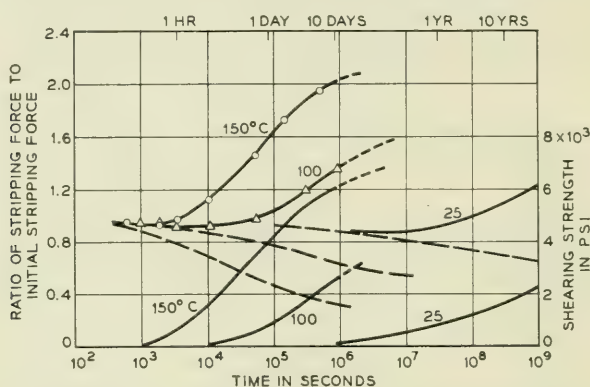


Fig. 14 — Stripping force and shear strength of bare copper on nickel silver connections as a function of time and temperature.

in greater mechanical strength and a faster fusion of the wire and terminal. Experiments show that there are at least four metals which diffuse faster than tin, but due to the economic problems and the brittleness of some of the alloys formed, none of them are being considered for solderless wrapped connection.

The reason the diffusion forces in the tin plated solderless wrapped connection do not cause an increase in strength of more than twenty per cent during the life of the connection is that the limiting shearing stress in tin is only 3,000 pounds per square inch. If now we substitute for tin plating a plating with a larger limiting shear stress, the strength of the connection should increase by a larger factor. To be of use, however, it is necessary that the diffusion forces shall develop rapidly.

The data of Fig. 10 show that if we can produce a higher shearing stress on a layer of aluminum, the activation energy can be lowered and

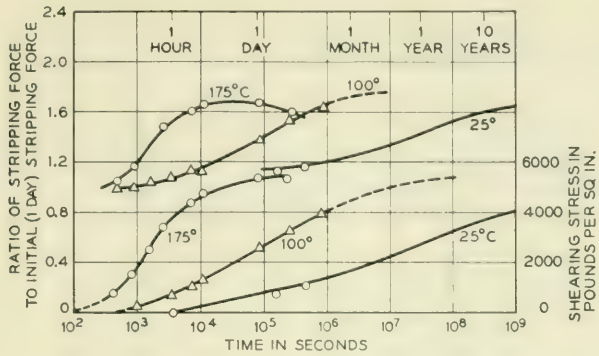


Fig. 15 — Stripping force and shear strength of copper-zinc-brass connections as a function of time and temperature.

be made to approach the cold welding condition. This requires that aluminum be placed on a stronger material such as brass, nickel silver or copper in order that the area of indentation for a given hoop stress will decrease and the shear stress will increase. Hence, the connection should form in a very short time. Furthermore, since the limiting shear stress for aluminum is near 6,000 pounds per square inch, one should expect that the strength of the connection will nearly double. Since aluminum is not easily electroplated this suggestion probably is not practical.

Of all the other metals examined, the next most promising are silver, cadmium and zinc. The activation energy for diffusion of zinc into copper

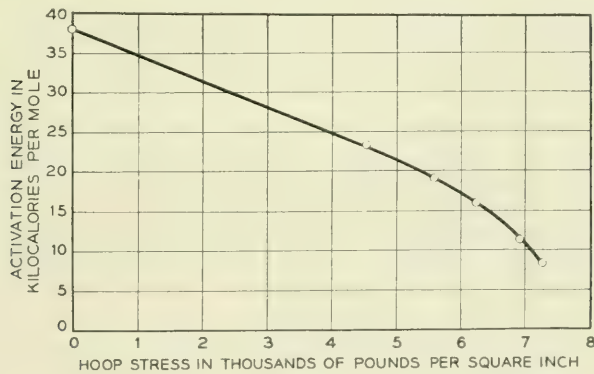


Fig. 16 — Activation energy for a copper-zinc-brass connection as a function of hoop stress.

at no stress is given⁶ as 38 kilocalories per mole which is 4 kilocalories higher than that for aluminum and 3 kilocalories under that for tin. The activation energy of silver and cadmium into copper also have low values. Hence, if the effect of stress parallels that for tinned copper, zinc and cadmium should diffuse into copper and nickel silver faster than tin. Furthermore, the limiting shearing 5000 pounds per square inch. Zinc, silver and cadmium are readily plated on the terminals or the wire.

Zinc plating has been tested experimentally by constructing a solderless wrapped connection of bare copper wire on zinc plated nickel silver or brass terminals. The stripping force for a bare copper wire wrapped on a terminal plated with 0.001 inch thickness of zinc has been measured at 175°C and at 100°C as a function of time with the results shown by Fig. 15. The strength increases to 60 per cent over that found initially in a time of less than two hours at 175°C. At 100°C, the time required is a little over a day. If we subtract the relaxed force from the stripping force and divide by the area of the connection, the shear strength of the connection is as shown by Fig. 15, lower curves, for the two temperatures. From these two curves, an activation energy versus hoop stress can be obtained with the results shown by Fig. 16. These values allow one to extend the time variation of the shear stress down to room temperature with the result shown by Fig. 15. Adding these values multiplied by the area of the connection to the relaxed hoop stress, the indicated stripping force at room temperature is shown by Fig. 15, room temperature curve. This force increases at such a fast rate that the strength can be observed to increase at room temperature and corresponding measurements are shown by the circles. Since corrosion cannot occur in a region of fusion, a criterion for the corrodability of a connection is the time required to complete half of the total fusion at room temperature. On this basis, the zinc plated connection has the lowest half fusion time of any of the materials tested.

Although zinc diffuses more readily than other materials examined, it tends to form more brittle alloys and hence its use has not been seriously considered for solderless wrapped connections.

⁶ R. M. Barrer, *Diffusion In and Through Solids*, Cambridge University Press, 1941, Table 67, p. 275.

A New Multicontact Relay for Telephone Switching Systems

By I. S. RAFUSE

(Manuscript received April 8, 1954)

The trend in new telephone switching systems toward faster operation, longer life and lower cost, indicates a need for faster and more capable control circuits. This paper describes a new high speed wire spring multicontact relay designed primarily for these applications. The basic unit contains 30 make-contact pairs. Two variations of the new design provide relays of 60-contact capacity. They are mechanically and electrically interchangeable with all crossbar system multicontact relays.

INTRODUCTION

In a modern dial telephone central office, many thousands of momentary intraoffice control connections are made daily between the various parts of the switching equipment. For example, in the No. 5 crossbar system, seven major types of connectors¹ are used to associate markers² with other common control circuits, and with the switching frames, for brief intervals, to assist in setting up the talking connection. Connectors are required to simultaneously close a large number of circuit paths, as many as 240 in the trunk link connector. The earlier flat spring type multicontact relays³ used for this purpose provide large blocks of contacts per relay and provide an economical means for common or multiple wiring.

The trend in new improved switching systems toward longer life, faster operation, lower cost and reduced maintenance, indicates a need for faster and more reliable connector circuits. This paper describes a new multicontact relay, designed primarily for these applications. It is a wire spring relay incorporating the improved manufacturing processes and many of the design features of the new general purpose wire spring relay.⁴ The basic unit, Fig. 1, for use in new equipments, is a high speed 30-contact pair relay, with wiring terminals arranged for horizontal multiple connections.

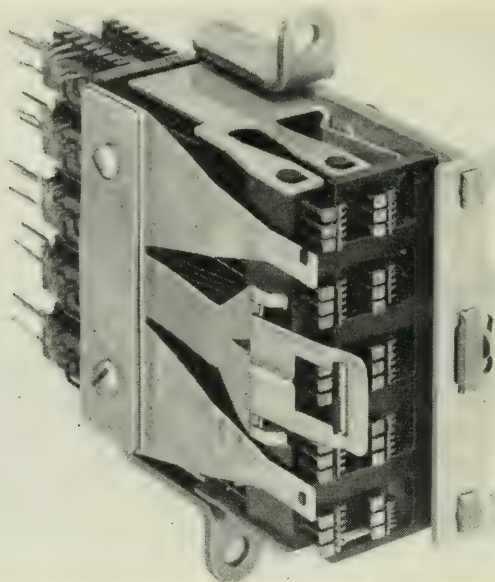


Fig. 1 — New 30-contact relay with contact cover detached.

This development also includes two modifications of the new wire spring relay for replacement use and for additions in existing crossbar equipments. They are 60-contact pair relays, each consisting of two 30-contact units attached to a common mounting bracket. They are completely interchangeable with existing crossbar system multicontact relays. When a new improved relay is developed, it is almost invariably necessary to continue in manufacture and carry in merchandise stock, small demand codes of the old relays for many years. In this case, however, manufacture of the old multicontact relays will be discontinued and all future needs will be supplied by the new product.

OBJECTIVES AND REQUIREMENTS

At the start of the project, all requirements from the standpoint of operating performance, circuit design and equipment use, were prepared in detailed form. The principal design objectives are summarized as follows:

Electrical and Mechanical

1. Operate and release times as fast as economically possible.
2. Forty year life, or 200 million operations, with no adjustment necessary during the first 100 million operations.

3. No contact chatter.
4. No false actuation due to armature rebound.
5. No magnetic or vibrational interference.
6. 120-ohm and 275-ohm coils, to work with equipment already in use.

Equipment

1. Lower manufacturing costs.
2. Reduced mounting space.
3. Terminal arrangement for multiple wiring same as at present, or equivalent from a wiring standpoint.

Maintenance

1. Contact failures due to dirt or insulating films should be substantially equal to and preferably less likely than in the present relay.
2. No contact locking due to contact erosion.
3. Contacts should be replaceable in the field.
4. Coil winding should be replaceable in the field.
5. Field adjustment should be reduced to a minimum.

Replacement Relay

The design objectives also included modification of the new relay, if possible, to replace multicontact relays in existing crossbar equipments.

Design History

During the early stages of this development, considerable effort was directed toward improving the present flat spring multicontact relay. Later, many experimental models were constructed, to investigate other flat spring, and wire spring designs, and several contact actuating methods. The most favorable designs of flat and wire spring multicontact relays were compared, and their differences were resolved by an analysis of manufacturing tolerances and their effect on performance and cost. Preliminary estimates of initial cost were only slightly in favor of the wire spring design. However, the wire spring relay offered significant advantages in (1) higher speed, longer life, less chatter, (2) better manufacturing control of tolerances, (3) less maintenance, and (4) possibilities of future cost reductions as further improvements are made in mechanized methods of manufacture.

DESCRIPTION OF THE NEW RELAY

General

Stationary single wire and moving twin wire spring subassemblies are arranged in alternate layers attached to the core and mounting

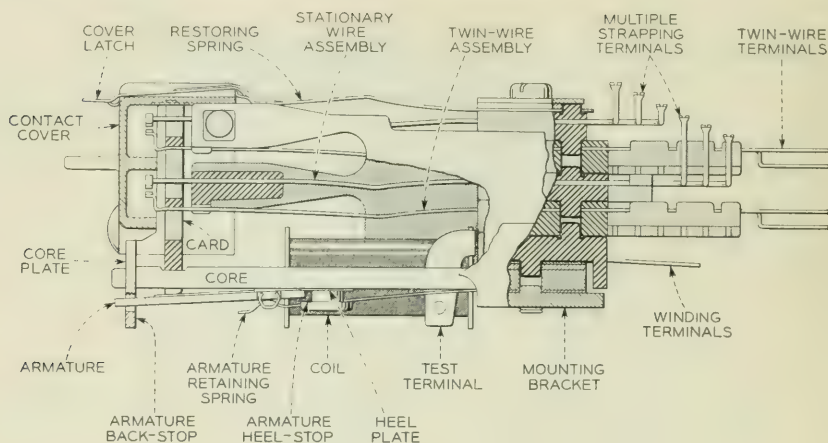


Fig. 2 — Top view of relay showing location of parts.

plate, as shown in Fig. 2. The wire spring assemblies form two rows, each containing 15 make contacts. Each contact pair consists of moving twin contacts on separate twin wires, associated with a single stationary contact. The stationary springs are supported close to the contacts by arms extending from the bracket. A detailed view of all parts and sub-assemblies for a 30-make contact relay is shown in Fig. 3. Moving contacts are pretensioned by relatively large pre-deflections as shown in Fig. 4. The method used in flat spring multicontact relays to obtain contact force is illustrated for comparison. It is apparent that contact force obtained by the "buckle" method depends on operating stud length and therefore is subject to change due to wear. The new pretensioned wire springs are supported and actuated by a single molded phenolic card by the "card release" method as illustrated in Fig. 5. In the unoperated position, the card is held against the core by a restoring spring, which also supplies the force to open all contacts. In the operated position, the armature supplies the force to move the card, releasing the twin wires and closing all contacts. This method of actuation has some important advantages:

1. Contact force is essentially independent of gauging and wear.
2. The effects of wear at points in the relay which affect gauging are compensating to some degree, and therefore tend to minimize changes in gauging.
3. Dimensional variations controlling contact separation and armature travel are reduced, making possible shorter armature travel, faster

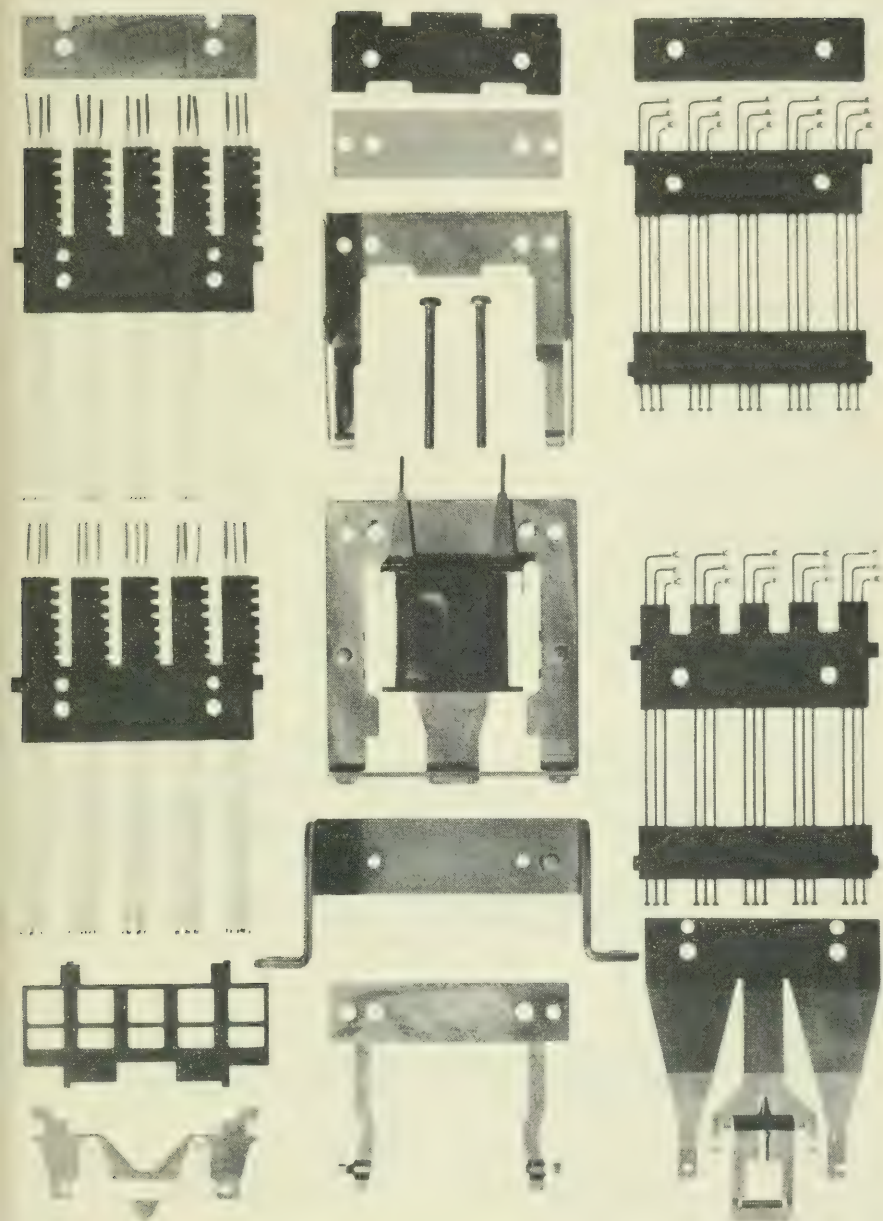


Fig. 3 — Parts of the new 30-contact relay.

operate and release times, less contact chatter and increased mechanical life.

Normally, the armature is held against the card by the lightly pre-tensioned armature hinge spring. However, when the relay is released, the armature motion becomes quite complex. Overtravel at the front is limited by the core plate backstop, and motion at the back, or heel sections, is limited by the heel stop studs. This freedom of movement is intentional, its purpose being to dissipate armature energy into the core plate and core rather than back into the card.

Magnetic Circuit and Armature

Analytical and experimental studies⁵ show that one per cent silicon iron, with its higher resistivity, relative freedom from aging, and lower eddy-current losses compared to ordinary magnetic iron, provides optimum speed in the new fast relay. The contact load, about twelve grams per twin contact pair, is about one-half the load required in the present relay, and this together with winding space and heating,⁶ largely determines the size of the magnet. The magnetic structure is shown in Fig. 6. The core is a one piece "E"-shaped section. The armature is a flat member made of low carbon steel having specific magnetic characteristics. This material simplifies manufacture, resulting in a cost saving with

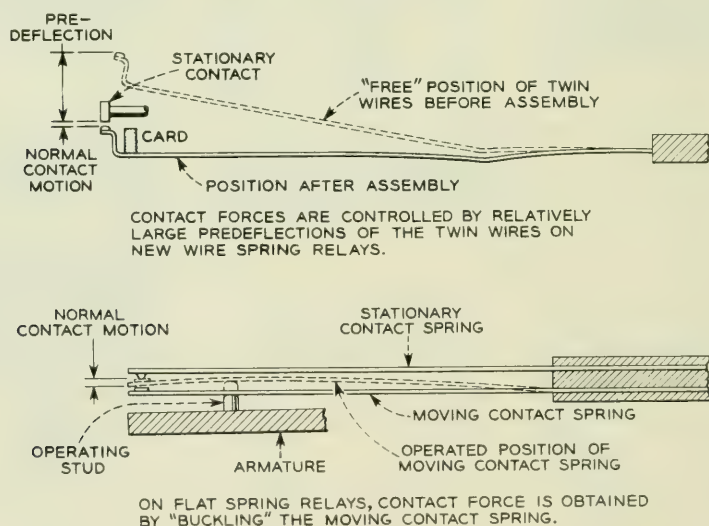


Fig. 4 — Development of contact forces in wire spring and flat spring multi-contact relays.

no appreciable penalty in performance. The armature has adequate sections to carry the flux, optimum poleface area and lowest possible mass. Two small rectangular holes in the armature locate the base of the card in the horizontal direction only. The card is located vertically by the restoring spring as illustrated in Fig. 1. Fast release is obtained by a nonmagnetic separator strip, welded to the face of the armature. This strip also provides a smooth supporting surface for the molded card. Negligible wear at this critical point contributes materially to long life and stable adjustment of this relay.

A cellulose acetate filled coil⁷ is assembled to the center leg of the core. A nonmagnetic core plate, illustrated in Fig. 7, is then forced over the three core legs to hold them in alignment. The center hole in the core plate also functions as an armature backstop and permits a certain amount of overtravel of the armature when the relay is released.

Coil Assemblies

For circuit reasons, 120-ohm and 275-ohm coil resistances used in the old relays are required in the new relays. Nominal power savings which ordinarily would result due to an improved magnet and reduced load, are therefore sacrificed in the new relays in favor of increased speed. More than half of the new relays are expected to be used in circuits requiring maximum speed of operation and will, therefore, have 120-ohm

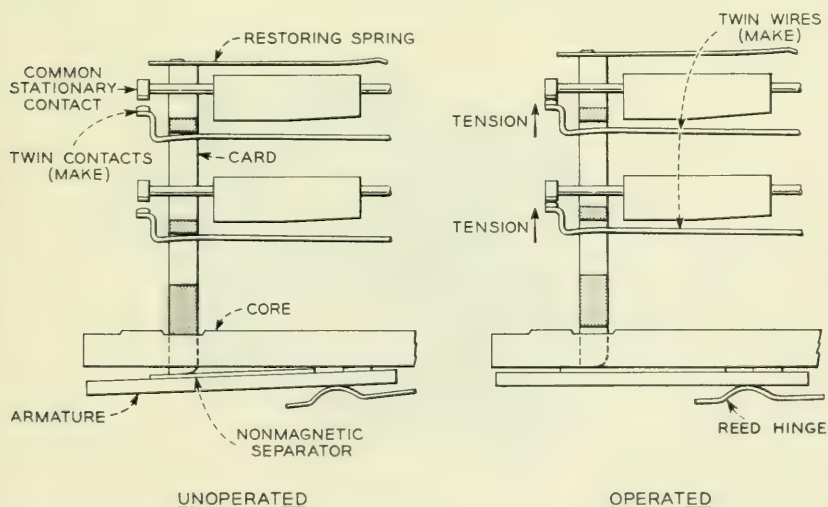


Fig. 5 — Principal of contact operation.

coils. In normal operation, these relays are not energized continuously, but operate only for short intervals while a talking circuit is established. The design provides for replacement of a coil winding in the field, but this is not expected to be necessary except in rare cases.

Molded Wire Spring Subassemblies

Wire spring subassemblies for a 30-contact relay are shown in Fig. 8. The two single wire subassemblies differ only in their terminal arrangement. The twin wire subassemblies are identical. Therefore only three basic molded parts are required which supply all needs for all new production relays. The wire spring sections are molded in continuous ladders⁸ as in the general purpose relay. Spring bending, contact welding and coining⁹ and terminal forming for solderless wrapped connections,¹⁰ are performed in automatic tools developed by Western Electric Company engineers. A comparison of the wire spring parts used in two new 30 make relays and the corresponding parts of a 60-make flat spring relay, is shown in Fig. 9. This illustrates the reduction in parts and simplicity of wire springs compared with flat springs. Seven types of subassemblies are used in the twelve layers required for an equivalent flat spring relay.

Terminals of the single wire subassemblies are formed for multiple wiring, and therefore differ in length and configuration. An improved

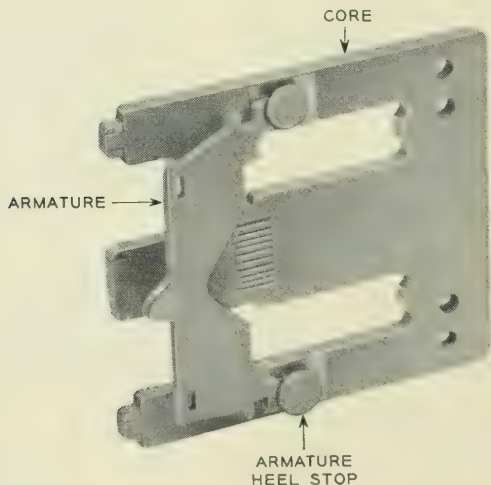


Fig. 6 — Magnetic structure of the new relays.

form of open wire multiple strapping has been developed for use with this terminal arrangement. It consists of bare wires held together in ladder form by means of phenolic plastic blocks molded successively in a continuous process. Fig. 10 shows relays with ladder type horizontal strapping soldered to the single wire terminals. The usual cable with new solderless wrapped connections is used on the twin wire terminals. Fig. 10 also illustrates the accessibility of multiple connections provided by locating them off to one side in the clear area between cable groups.

Contacts

All contacts in the new wire spring relays are palladium having a volume of contact material suitable for forty years life. This is equivalent to about 200 million operations for relays in high usage circuits. Contacts are easily visible, readily cleaned and may be insulated for test purposes.

Contacts may be replaced in the field if necessary using Bell System field welding equipment.¹¹ Suitable tools and electrodes have been developed to permit use of this equipment on all wire spring relays.

Assembly and Adjustment

The relay pile-up is securely fastened by two high tensile screws and a spring compliance member. Laboratory tests show that a pile-up of

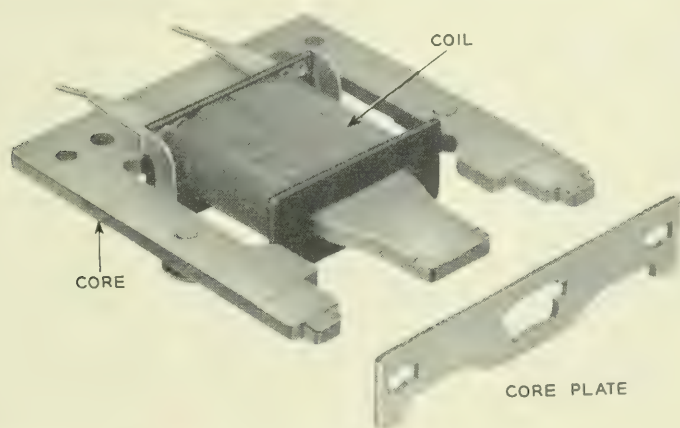


Fig. 7 - Core legs are held in alignment by the core plate, which is forced on the ends after the coil is assembled.

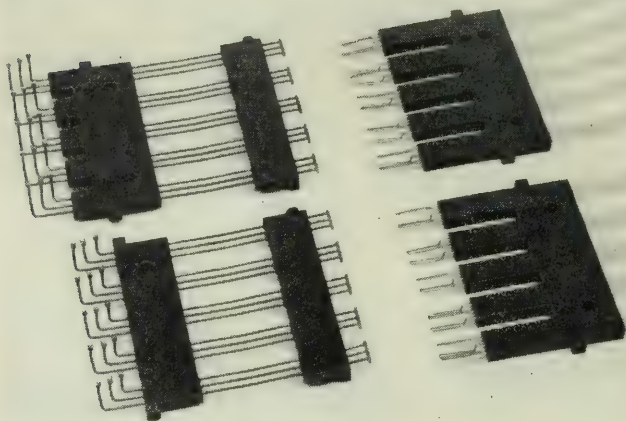


Fig. 8 — Molded wire assemblies for a 30-contact relay.

this design will remain tight under widely varying atmospheric conditions.

In the design of the new relay considerable attention was given to manufacturing control of tolerances, with reduced assembly and adjustment costs as an objective. The molding process provides dowels for aligning the four layers of contact springs, trunnion supports to locate the single wires, and control grooves in the single wire subassemblies to align the twin wires. These features provide accurate registration for mating contacts, for the location of wire subassemblies relative to other relay parts and for the pretensioned restoring spring which in turn locates the actuating card in relation to the operating contacts. The card determines contact separation of all twin moving contacts in relation to their respective single stationary contacts. Due to these controls, and because the pre-deflected twin wire springs require no adjustment of contact force, no factory adjustments are anticipated except on relays which fall outside acceptable limits for back tension or contact gauging (or follow) as assembled. If necessary, therefore, the restoring spring may be adjusted to control the armature back tension. Contact gauging may be controlled if required by independent mass adjustment of the single wire contact rows. Fig. 11 shows this operation being performed by bending the bracket arm using a tool designed for this purpose. The mass adjustment feature is expected to simplify field maintenance practice.

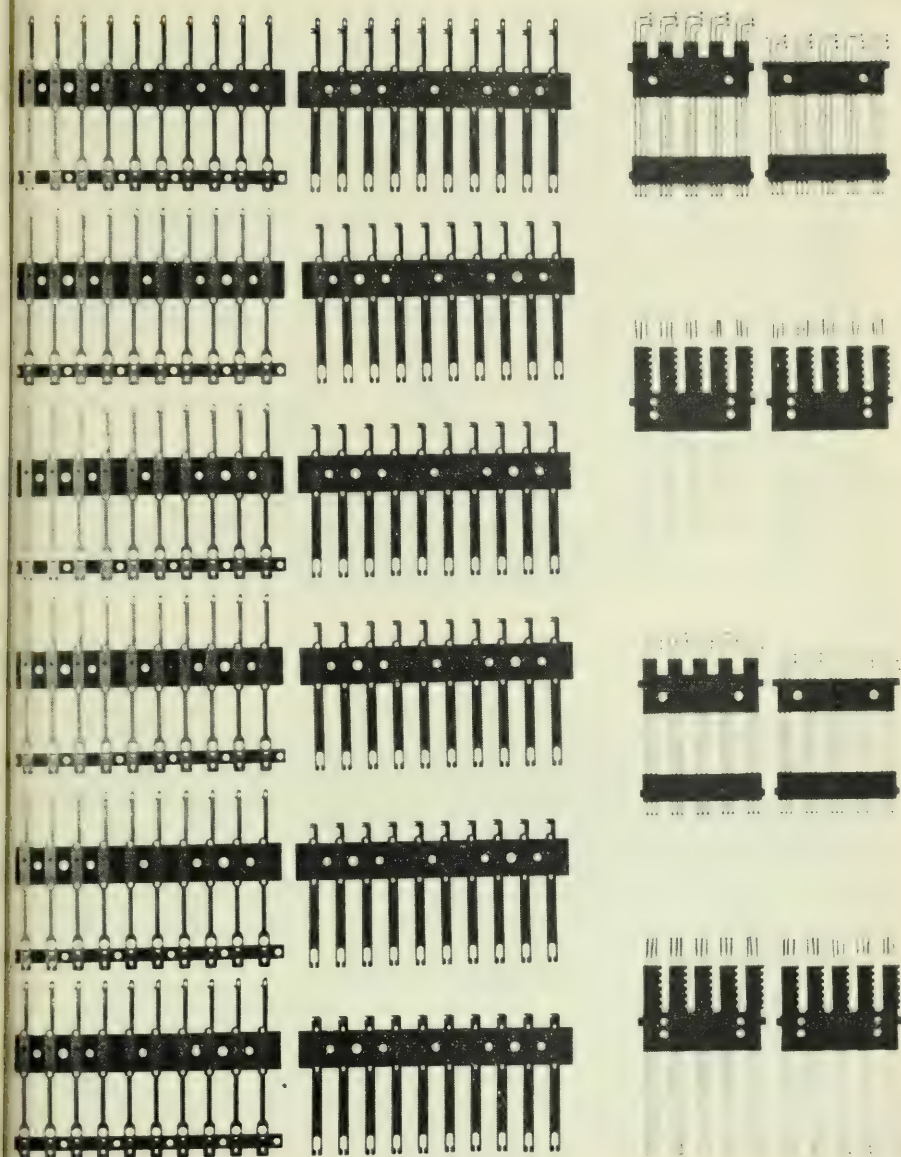


Fig. 9 — A comparison of molded wire assemblies for a 60-contact replacement relay with the corresponding flat spring relay parts.

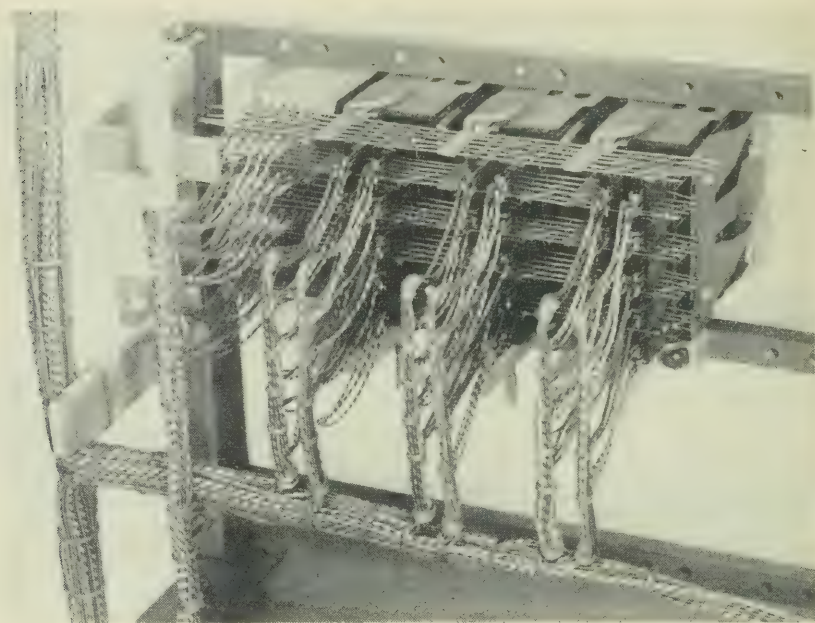


Fig. 10 — New 30-contact relays showing multiple wiring by the new ladder type strapping method, and the new solderless wrapped cable connections.

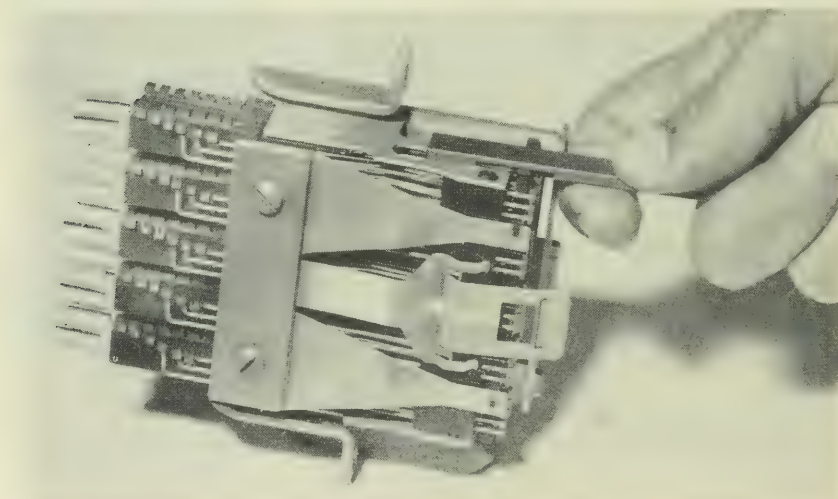


Fig. 11 — New 30-contact relay showing method of mass adjusting stationary contacts.

Replacement Type Relays

Future replacements of flat spring multicontact relays in existing crossbar equipments for maintenance reasons require a slightly modified form of the new relay, capable of complete interchangeability. As these relays are commonly used in connector circuits, operate and release times must be comparable with older type relays in order to avoid circuit interference. This was accomplished in the new wire spring relay with only minor changes in design and in the manufacturing process.

As illustrated by Fig. 12, the interchangeable or replacement relay consists essentially of two 30-contact units assembled on a common mounting bracket having the same vertical mounting centers as the 60-contact flat spring relay. Since horizontal mounting space required by the new relay has been reduced, the 60-contact wire spring unit may also be used to replace 30-, 40- and 50-contact flat spring multicontact relays. The model shown in Fig. 12 has terminals arranged for horizontal multiple wiring. In another variation of the replacement relay all terminals are arranged for cable wiring.

Modifications necessary to slow down the new relays for replacement use are (1) longer armature travel, requiring a different card; (2) an armature of larger mass — although this is a different piece part it has the same contour as the armature for the fast relay and may be punched in the same tool setup; (3) a core of low carbon steel in place of the one per cent silicon iron used in the high speed relay; (4) a leakage reluctance shunt element shown in Fig. 13 to by-pass a portion of the total magnetic flux; and (5) coil windings having resistances of 120 and 275 ohms as required for circuit reasons, but having the number of turns calculated for slower speed consistent with reliable operating capability.¹²

Relay Performance

Measurements have been made on laboratory-built models carefully prepared and adjusted to simulate extreme ranges of manufacturing tolerances, and more recently, on representative samples of pre-production relays. Although some of the performance characteristics studied will be determined accurately only after long-term use in the field, it has been possible by designed experiments and comparative tests, to obtain a fair appraisal of relay capability. These tests and measurements indicate that design objectives stated earlier in this paper have been substantially achieved.

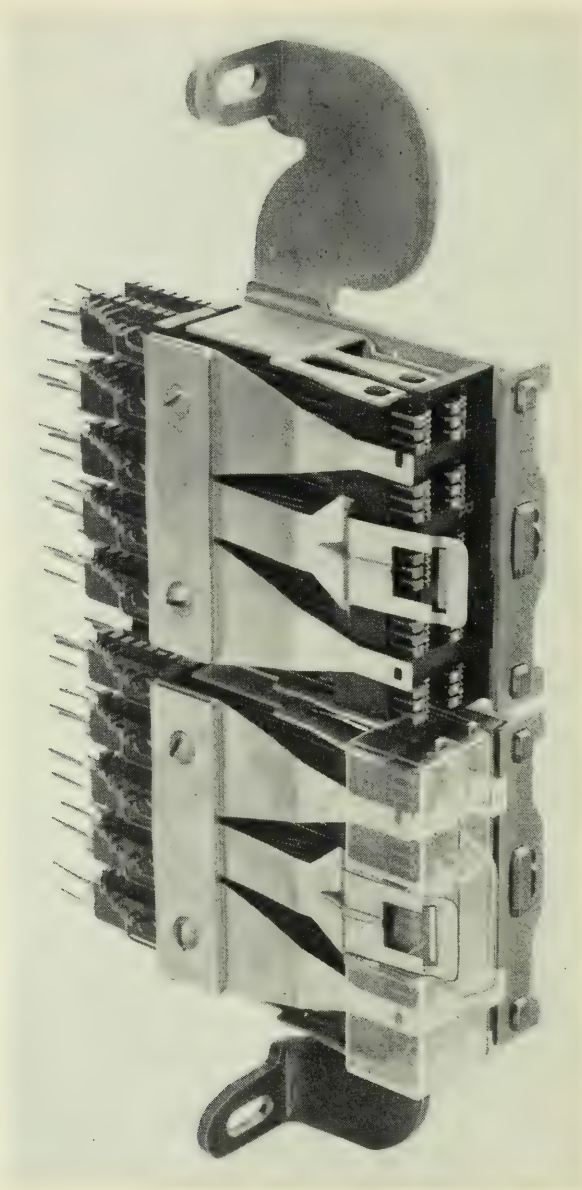


Fig. 12 — A 60-contact replacement relay with one contact cover detached.

Load and Pull Characteristics

Load and pull curves are measured under essentially static conditions. Spring load and armature motion are both observed at the center line of the card. They are measured and simultaneously recorded in chart form in a modification of a tensile testing machine.¹³ A typical load and pull chart is shown in Fig. 14. The abscissa shows armature motion, as the armature moves the card and the spring load, through a distance of 0.030 inch to the operated position, and back again. The ordinate shows spring load on the armature on both operate and release, and also the magnetic pull which is developed in the armature for various numbers of ampere turns in the winding.

The load and pull chart provides a comprehensive picture of over-all relay performance. For example, starting from the released position, the force or back tension, holding the card against the core is about 140 grams. Following the upper curve, the spring load increases slowly as the armature moves toward the core, until the first contacts make at a load of about 200 grams. The load increases rapidly as the remaining contacts are closed until the last contacts are closed at about 650 grams. Further travel of the armature to the operated position increases the spring load to a final value of about 700 grams. As the armature is allowed to return to the original position, the lower curve is traced. The area between the two curves is a measure of mechanical hysteresis, or friction, in the relay. This energy loss is a very small fraction of the spring load at all values of armature travel.

The pull curves show ampere turns necessary to assure operation of

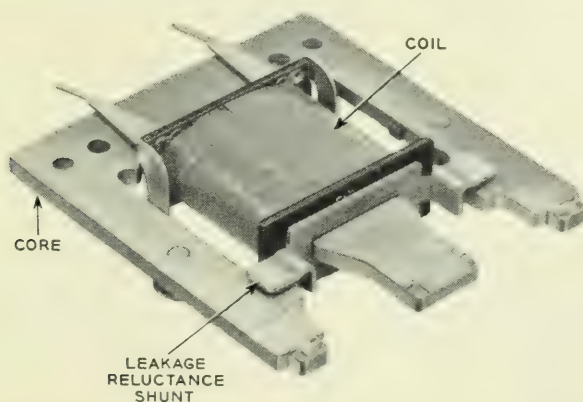


Fig. 13 Core assembly for replacement type relays showing leakage reluctance shunt.

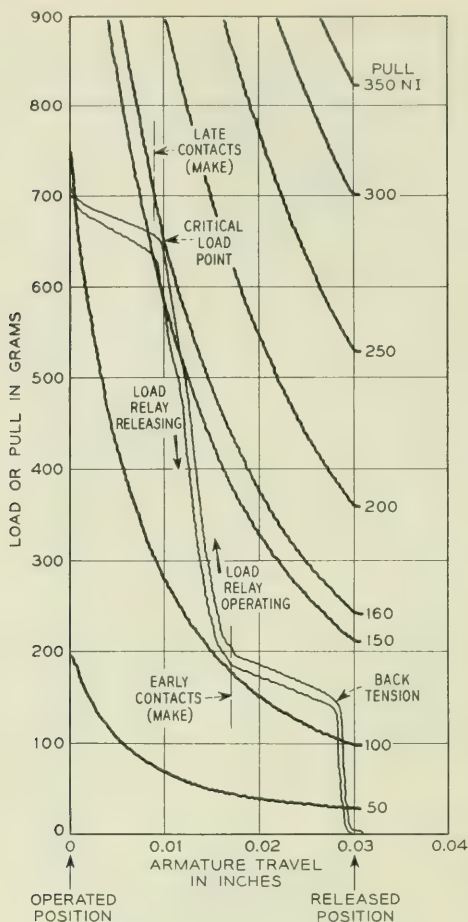


Fig. 14 — Typical load and pull characteristics of a 30-contact fast relay.

the relay. The maximum ampere turns required are determined by the "critical load point." This occurs at 0.010 inch armature travel and about 650 grams. Under static conditions, therefore, 160 ampere turns would be required for complete operation. Circuit uses for these relays do not include nonoperate, hold, or release requirements. This information could however be obtained from the pull curves in a similar manner.

Operate and Release Speed

The new high speed multicontact relay operates two to three times as fast as its predecessor, the flat spring type relay. Operate and release

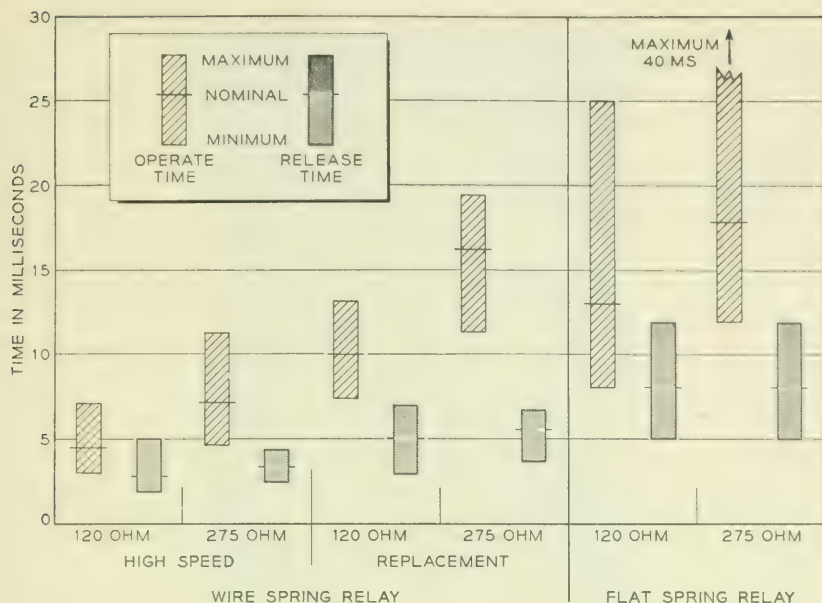


Fig. 15 — Comparison of operate and release times of wire spring versus flat spring multicontact relays.

times are shown in Fig. 15. The improved performance of the new relays is shown by nominal operate and release values, and also by greatly reduced spread, or difference between minimum and maximum values, as compared with corresponding data for flat spring relays.

Minimum operate times of the replacement and existing flat spring relays are comparable. It will be noted however, that operate and release time spreads are much less in the new relay. This generally should improve the operation of existing crossbar circuits as new relays are used for replacements or additions.

Contact Performance

As speed increases, relays and other switching mechanisms become more susceptible to false operation of contacts. It is obvious also that faster operation adds to life requirements and therefore extends the period or increases the number of operations during which trouble-free contact performance must be provided. For these reasons, extensive studies were made of chatter, unprotected erosion, and locked contacts as applied to the new multicontact relay. Additional longer range tests

are planned to study protected contact erosion and susceptibility to open contact failures.

When contacts of the new high speed relay are operated, initial chatter does not normally exceed 0.1 millisecond. There is no shock chatter caused by spring vibration due to impact of the armature with the core or backstop. There is no chatter caused by hesitation of the armature in its travel at the point where it picks up the contact load. This is due largely to the low mass of 0.020 inch diameter twin wires, the low mass and short travel of the actuating mechanism, the type of card operation, and the rigid mounting of the relay structure.

Electrical erosion of contact material is reduced in the new relay, because of the reduction in chatter. For this reason, the contact size provided is expected to be adequate for all normal use for the life of the relay, and contact maintenance should be greatly reduced.

Locked contacts are substantially eliminated in the new relay by the single card release method of operation. Static and dynamic forces associated with the restoring spring and card system are powerful enough to break loose any random pair of locked contacts.

Contact failures due to dirt or the formation of insulating films on the contacts are difficult to check in laboratory tests. Long-term accelerated tests are necessary, with a large test sample, under carefully controlled dust conditions. Many precautions were taken in the design to minimize failures from this source, as follows: (1) a dust cover, Fig. 16, encloses the contacts, but does not enclose the coil; (2) the cover partially segregates the contacts in groups of three pairs of contacts, reducing air movement in the vicinity of the contacts; (3) palladium contact material is used on all relays; (4) twin contacts are coined — the rounded surface reduces the area in contact, effectively restricts the area which may trap lint or other foreign matter, and increases contact pressure; (5) card release actuation and wire springs with large pre-deflections insure that no appreciable loss of contact force will occur due to age or erosion; and (6) twin contacts are attached to completely independent wires.

Rebound chatter is another form of false operation which occurs in the form of contact reclosures caused by rebound of the armature after striking the backstop when the relay is released. Fundamental studies were made of rebound behavior in relay structures¹⁴ and various models were constructed and measured. As a result there is no rebound chatter in the new high-speed wire spring relay within the range of normal adjustment. A comprehensive survey was made to determine the probability of reclosures due to rebound, in relays having limiting adjust-

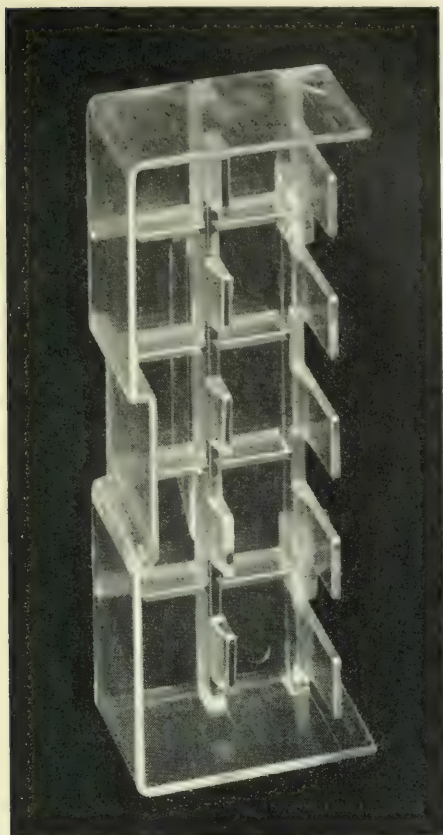


Fig. 16 — Contact cover for the new relays showing compartments in which contacts are grouped.

ments. The probability of reclosures exceeding one millisecond duration is estimated as one in 28,000 relays. This grade of performance is due primarily to: (1) an armature having low travel and the lowest possible mass; (2) an armature suspension designed to dissipate rebound energy into the core plate and core rather than into the actuating card; and (3) a stiff mounting bracket to reduce the natural amplitude of core vibration due to armature impact on operate and on release.

Life

Less than 5 per cent of the new multicontact relays will be required to operate more than 100 million times in crossbar systems during an esti-

mated life term of forty years. Life tests show that no readjustment should be necessary during the first 100 million operations. The tests also indicate that not more than a small percentage of relays will require readjustment prior to an estimated maximum life of 200 million operations. In extreme cases, still greater life may be obtained if required, by replacing the molded card. This is an inexpensive part and replacement is easily accomplished.

Stability

When the new relays are exposed to extreme temperature and humidity cycles, the greatest change in contact separation is, in general, about 0.004 inch, and only a small percentage of relays are likely to be used in this manner. Tests indicate that changes of this magnitude leave adequate margin for 100 million operations before readjustment is necessary.

For economy, most equipment is shop assembled and wired on a frame basis, and shipped complete, ready for installation as equipment units. It is important, therefore, that apparatus units should be capable of withstanding physical shock far in excess of normal usage. Design features in the new relay which provide an adequate margin of safety in this respect are: (1) a rigid mounting bracket; (2) the wire spring pile-up is attached securely to the bracket with two specially heat treated steel screws; (3) the cover is held in place by the bending moment of an embossed section of a spring clip with a force many times greater than the compressive force of a single spring; and (4) guard surfaces molded in the cover prevent twin wires from leaving their respective guide notches in the single wire combs.

Excessive shocks during shipment have, at times, damaged flat spring relays by bending their brackets. The new relays have been subjected to shocks of similar magnitude without damage.

Magnetic Interference

Under certain marginal conditions, a relay may be affected by leakage flux from adjacent relays entering its magnetic circuit, and changing its operate and release values. Tests show that interaction is negligible between the new relays and also between new and old type relays when they are used in adjacent positions.

CONCLUSIONS

The new 30-contact relays provide faster operate and release times, longer life, improved contact performance, reduced maintenance, and

greater adaptability in new circuit and equipment units than previous multicontact relays. The new relays also require less vertical and horizontal space in new equipments. As a result of these improvements, substantial savings are expected when these relays are used. The design includes many features which permit the use of mechanized manufacturing processes, which, in turn provide better control of tolerances. For these reasons, lower initial costs are expected as manufacturing and assembly methods continue to improve.

The new 60-contact relays are completely interchangeable with all codes of flat spring multicontact relays in existing crossbar equipments, and in addition, they provide superior performance, longer life and reduced maintenance. Therefore, manufacture of the flat spring multicontact relays will be discontinued as soon as new relay production becomes adequate for all uses.

ACKNOWLEDGEMENTS

Many design problems required the cooperation, special knowledge and facilities of Materials, Chemical and Research Departments of Bell Telephone Laboratories. Some design features, particularly those involving new manufacturing processes, were developed in close cooperation with Western Electric Company engineers.

These acknowledgements would not be complete without including the technical contributions and assistance of the many people in Switching Apparatus Development Department who were directly and indirectly associated with the project, and to E. G. Walsh and T. H. Guetlich for their assistance in the preparation of material and illustrations for this paper.

REFERENCES

1. G. S. Bishop, Connectors for the No. 5 Crossbar System, Bell Labs. Record, **28**, p. 56, Feb., 1950.
2. A. O. Adam, The No. 5 Crossbar Marker, Bell Labs. Record, **28**, p. 502, Nov. 1950.
3. Bruce Freile, Multicontact Relay, Bell Labs. Record, **17**, p. 301, May, 1939.
4. A. C. Keller, A New General Purpose Relay for Telephone Switching Systems, B. S. T. J., **31**, p. 1023, Nov., 1952.
5. R. L. Peek, Jr. and H. N. Wagar, Magnetic Design of Relays, B. S. T. J., **33**, p. 23, Jan., 1954.
6. R. L. Peek, Jr., Internal Temperatures of Relay Windings, B. S. T. J., **30**, p. 141, Jan., 1951.
7. C. Schneider, Cellulose Acetate Filled Coils, Bell Labs. Record, **29**, p. 514, Nov., 1951.
8. A. J. Brunner, H. E. Cosson and R. W. Strickland, Wire Straightening and Molding for Wire Spring Relays, B. S. T. J., **33**, p. 859, July, 1954.
9. A. L. Quinkan, Automatic Contact Welding in Wire Spring Relay Manufacture, B. S. T. J., **33**, p. 897, July, 1954.

10. Solderless Wrapped Connections, B.S.T.J., **32**, May, 1953. Introduction, J. W. McRae, pp. 523 and 524. Part I — Structure and Tools, R. F. Mallina, pp. 525-556. Part II — Necessary Conditions for Obtaining a Permanent Connection, W. P. Mason and T. F. Osmer, pp. 557-590. Part III — Evaluation and Performance Tests, R. H. Van Horn, pp. 591-610.
11. W. T. Prichard, Relay Contact Welder, Bell Labs. Record, **22**, p. 374, Apr., 1944.
12. M. A. Logan, Design of Optimum Windings, B. S. T. J., **33**, p. 114, Jan., 1954.
13. E. G. Walsh, Continuously Recorded Relay Measurements, Bell Labs. Record, **32**, p. 27, Jan., 1954 and H. N. Wagar, Relay Measuring Equipment, B. S. T. J., **33**, p. 3, Jan., 1954.
14. E. E. Sumner, Relay Armature Rebound Analysis, B. S. T. J., **31**, p. 172, Jan., 1952.

Topics in Guided Wave Propagation Through Gyromagnetic Media

Part III — Perturbation Theory and Miscellaneous Results

By H. SUHL and L. R. WALKER

Some problems, complete discussion of which would be extremely difficult, are treated approximately by means of perturbation theory. Among these are the partially filled cylindrical waveguide, and the problem of multiple internal reflections in a sample of finite length filling the cross section of a cylindrical guide. Propagation in a ferrite between parallel planes, magnetized along the propagation direction is discussed by the methods described in Part I. The paper concludes with an addendum to Part I — a numerical study of field patterns of the TE_{11} -limit and TM_{11} -limit mode for various dc magnetic fields.

INTRODUCTION

Parts I and II of this paper were devoted to a number of specific propagation problems, whose solutions, though frequently quite complicated, could be discussed with a reasonably modest investment of effort. Unfortunately, not all of these problems pertain to situations met with in actual gyromagnetic devices. Actual devices frequently employ structures whose performance could be predicted only as the result of lengthy computing programs. For example, the microwave gyrator using Faraday rotation usually employs a ferrite sample whose cross-section only partly fills that of the cylindrical waveguide. Although it is easy to formulate the corresponding equation for the propagation constant, the classification and survey, let alone the computation of solutions, would be very difficult to carry out.

Thus, one must often be content with approximate results, and the bulk of the present paper is devoted to perturbation methods. These take as starting point a situation whose propagation problem is essentially solved. The small change in propagation constant due to a slight change in the original state of the system is then calculated. The small

change of state may be the weak magnetization of an originally unmagnetized specimen occupying a substantial part of the structure, or the introduction of a very small specimen with arbitrarily large magnetization into the originally empty structure. Under the heading, "Small Magnetization — Arbitrary Sample-Size," we shall discuss the propagation constant for a pencil of ferrite of any radius, coaxial with a cylindrical waveguide, the space between guide wall and pencil being filled with an isotropic medium whose dielectric constant equals that of the ferrite. This is discussed in preparation for the practically more important case of a ferrite pencil of any radius in an air-filled guide. Here the unperturbed state of the system, when the pencil is unmagnetized and therefore isotropic, is already rather complicated and require some preliminary calculations. Under the heading "Small Sample-Size — Arbitrary Magnetization," we consider the case of a thin pencil of ferrite in an originally empty guide.

Another topic, not easily treated except by perturbation methods, is that concerning end effects in samples of finite length. After a preliminary discussion of internal reflections in an extended slab of ferrite (a problem which can be treated rigorously), two cases are considered: a ferrite slug of arbitrary length, closely fitting a cylindrical guide, and a thin disc normal to the guide axis, of arbitrary size. In these cases interest centers around the effect of sample length on Faraday rotation, though for the ferrite slug a subsidiary effect, that of mode conversion, is also mentioned briefly.

It should be emphasized that the perturbation methods employed here are not in themselves novel. They are standard to most linear eigenvalue problems of physics, and have been used in connection with electromagnetic problems by many authors.¹

The remainder of the paper is devoted to a discussion of a ferrite-filled "cable" in plane parallel form, using the methods of Part I. The treatment is kept in terms of saturation magnetization and magnetizing field, and is based on Polder's equations. The paper concludes with an addendum to Part I, which reports some calculations and graphs of field patterns in a cylindrical waveguide completely filled with ferrite.

1. PERTURBATION METHODS

1.1 *General Method*

A number of authors have made applications of perturbation theory to the problems of propagation in gyromagnetic media and the exposition which follows is included mainly for completeness. We shall develop the subject in the following fashion: it will be supposed that the unper-

turbed system is a wave guide containing a medium whose permeability and dielectric tensors are diagonal and isotropic, but may vary over the cross section of the guide, although not in the z -direction along the guide. For this system it will be assumed that a complete set of normal modes exists for which appropriate orthogonality relations are known. The perturbation of the system will then consist of changes in the permeability and dielectric tensors of the medium, including the addition of non-diagonal terms. If these changes are to be genuine perturbations, they must be of one of two kinds. Either, the variation in the properties of the medium is confined to a limited region, small in volume in some appropriate sense, in which case its magnitude may be large, or, we may have a small fractional change in the material properties extending over a considerable volume. The fields in the guide may be expanded in the normal modes and a system of equations is developed for the z -dependence of the amplitudes of these modes. These equations are then solved approximately, making use of the smallness of the perturbing terms. The results may then be specialized to the various situations of interest.

Let us suppose that the unperturbed permeability and dielectric constant are $\mu_1(x, y)$ and $\epsilon_1(x, y)$ respectively and that the system is now altered so that it possesses a permeability tensor

$$\begin{vmatrix} \mu_2(x, y) & -j\kappa(x, y) & 0 \\ j\kappa(x, y) & \mu_2(x, y) & 0 \\ 0 & 0 & \mu_3(x, y) \end{vmatrix}$$

and a dielectric tensor

$$\begin{vmatrix} \epsilon_2(x, y) & -j\eta(x, y) & 0 \\ j\eta(x, y) & \epsilon_2(x, y) & 0 \\ 0 & 0 & \epsilon_3(x, y) \end{vmatrix}.$$

Maxwell's equations for the perturbed system may be written, using the notation of Parts I and II,[†] in the form:

$$\begin{aligned} \nabla^* H_z - \frac{\partial H_t^*}{\partial z} - j\omega\epsilon_2 E_t - \omega\eta E_t^* &= 0, \\ \nabla^* E_z - \frac{\partial E_t^*}{\partial z} + j\omega\mu_2 H_t + \omega\kappa H_t^* &= 0, \\ \nabla \cdot H_t^* - j\omega\epsilon_3 E_z &= 0, \\ \nabla \cdot E_t^* + j\omega\mu_3 H_z &= 0. \end{aligned} \quad (1)$$

[†] We omit the vector signs from all transverse vectors, which are sufficiently labelled by the subscript " t ."

These may be rearranged as

$$\begin{aligned}\nabla^* H_z - \frac{\partial H_t^*}{\partial z} - j\omega\epsilon_1 E_t &= j\omega(\epsilon_2 - \epsilon_1)E_t + \omega\eta E_t^* = A_t, \\ \nabla^* E_z - \frac{\partial E_t^*}{\partial z} + j\omega\mu_1 H_t &= -j\omega(\mu_2 - \mu_1)H_t - \omega\kappa H_t^* = B_t, \quad (2) \\ \nabla \cdot H_t^* - j\omega\epsilon_1 E_z &= j\omega(\epsilon_3 - \epsilon_1)E_z = A_z, \\ \nabla \cdot E_t^* + j\omega\mu_1 H_z &= -j\omega(\mu_3 - \mu_1)H_z = B_z,\end{aligned}$$

where A_t , B_t , A_z and B_z are introduced as abbreviations for the terms on the right hand side of the equations. E_z and H_z may be eliminated by substituting in the first two equations the expressions

$$E_z = \frac{\nabla \cdot H_t^* - A_z}{j\omega\epsilon_1}$$

and

$$H_z = \frac{-\nabla \cdot E_t^* + B_z}{j\omega\mu_1}.$$

The two equations so obtained are

$$\frac{1}{j\omega} \nabla^* \left(\frac{\nabla \cdot E_t^*}{\mu_1} \right) + \frac{\partial H_t^*}{\partial z} + j\omega\epsilon_1 E_t = -A_t + \frac{1}{j\omega} \nabla^* \frac{B_z}{\mu_1}$$

and

$$\frac{1}{j\omega} \nabla^* \left(\frac{\nabla \cdot H_t^*}{\epsilon_1} \right) - \frac{\partial E_t^*}{\partial z} + j\omega\mu_1 H_t = B_t + \frac{1}{j\omega} \nabla^* \frac{A_z}{\epsilon_1}. \quad (3)$$

We now suppose that E_t and H_t can be expanded in the form

$$E_t = \sum_n a_n(z) E_{tn}(x, y)$$

and

$$H_t = \sum_n b_n(z) H_{tn}(x, y),$$

where

$$E_{tn} e^{-j\beta_n z} \quad \text{and} \quad H_{tn} e^{-j\beta_n z}$$

satisfy the unperturbed form of equations (3) and the boundary condition that tangential E vanishes at the guide wall. These equations have solutions for certain values of β_n only, but, if E_{tn} , H_{tn} are solutions for $\beta_n = c > 0$, then E_{tn} , $-H_{tn}$ are solutions for $\beta_n = -c$. We shall assume

hereafter that E_{tn} and H_{tn} pertain to positive β_n values. For a given unperturbed mode it follows that $\frac{b_n(z)}{a_n(z)}$ reverses sign when the direction of propagation reverses. Substituting these series for E_t and H_t in equations (3), one finds

$$\sum_n \left[\frac{db_n}{dz} + j\beta_n a_n \right] H_{tn}^* = -A_t + \frac{1}{j\omega} \nabla^* \frac{B_z}{\mu_1} \quad (4a)$$

and

$$\sum_n \left[\frac{da_n}{dz} + j\beta_n b_n \right] E_{tn}^* = -B_t - \frac{1}{j\omega} \nabla^* \frac{A_z}{\epsilon_1}. \quad (4b)$$

The orthogonality relation between functions of different n has been given by Adler.² It is

$$\int E_{tn}^* \cdot \tilde{H}_{tm} dS = 0,$$

where $n \neq m$; the tilde denotes the complex conjugate and the integration goes over the cross section of the guide. We consider the fields to be un-normalized and write

$$\int E_{tn}^* \cdot \tilde{H}_{tn} dS = \Delta_n,$$

Clearly

$$\int H_{tn}^* \cdot \tilde{E}_{tn} dS = -\bar{\Delta}_n. \quad (5)$$

Multiplying equation (4b) by $\tilde{H}_{tn}$ and equation (4a) by $\tilde{E}_{tn}$ and integrating each expression over the cross section we have

$$\begin{aligned} -\bar{\Delta}_n \left[\frac{db_n}{dz} + j\beta_n a_n \right] &= -\int A_t \cdot \tilde{E}_{tn} dS + \frac{1}{j\omega} \int \tilde{E}_{tn} \cdot \nabla^* \frac{B_z}{\mu_1} dS, \\ \Delta_n \left[\frac{da_n}{dz} + j\beta_n b_n \right] &= -\int B_t \cdot \tilde{H}_{tn} dS - \frac{1}{j\omega} \int \tilde{H}_{tn} \cdot \nabla^* \frac{A_z}{\epsilon_1} dS. \end{aligned}$$

Now we use the identity

$$\int_{\text{surface}} G_t \cdot \nabla^* F dS = - \int_{\text{boundary}} F (G_t \cdot ds) + \int_{\text{surface}} F \nabla \cdot G^* dS.$$

This yields

$$\begin{aligned}\int \tilde{E}_{tn} \cdot \nabla^* \frac{B_z}{\mu_1} dS &= - \int_{b'dry} \frac{B_z}{\mu_1} (\tilde{E}_{tn} \cdot ds) + \int \frac{B_z}{\mu_1} \nabla \cdot \tilde{E}_{tn}^* dS, \\ &= j\omega \int B_z \tilde{H}_{zn} dS,\end{aligned}$$

since $E_{tn} \cdot ds = 0$. H_{zn} is the H_z field appropriate to the n^{th} mode. Again, we have

$$\begin{aligned}\int \tilde{H}_{tn} \cdot \nabla^* \frac{A_z}{\epsilon_1} &= - \int_{b'dry} \frac{A_z}{\epsilon_1} (\tilde{H}_{tn} \cdot ds) + \int \frac{A_z}{\epsilon_1} \nabla \cdot \tilde{H}_{tn}^* dS, \\ &= -j\omega \int A_z \tilde{E}_{zn} dS.\end{aligned}$$

since $A_z = 0$ on the boundary. E_{zn} is now the E_z field of the n^{th} mode. Making use of these results we obtain the equations for the amplitudes in the form:

$$\begin{aligned}\frac{db_n}{dz} + j\beta_n a_n &= \frac{1}{\Delta_n} \left[\int A_t \cdot \tilde{E}_{tn} dS - \int B_z \tilde{H}_{zn} dS \right], \\ \frac{da_n}{dz} + j\beta_n b_n &= \frac{1}{\Delta_n} \left[- \int B_t \cdot \tilde{H}_{tn} dS + \int A_z \tilde{E}_{zn} dS \right].\end{aligned}\tag{6}$$

Restoring the full expressions for A_t , B_t , A_z and B_z , we have

$$\begin{aligned}\frac{db_n}{dz} + j\beta_n a_n &= \frac{j\omega}{\Delta_n} \left[\int (\epsilon_2 - \epsilon_1) E_t \cdot \tilde{E}_{tn} dS - j \int \eta E_t^* \cdot \tilde{E}_{tn} dS \right. \\ &\quad \left. + \int (\mu_3 - \mu_1) H_z \tilde{H}_{zn} dS \right],\end{aligned}\tag{7a}$$

$$\begin{aligned}\frac{da_n}{dz} + j\beta_n b_n &= \frac{j\omega}{\Delta_n} \left[\int (\mu_2 - \mu_1) H_t \cdot \tilde{H}_{tn} dS - j \int \kappa H_t^* \cdot \tilde{H}_{tn} dS \right. \\ &\quad \left. + \int (\epsilon_3 - \epsilon_1) E_z \tilde{E}_{zn} dS \right].\end{aligned}\tag{7b}$$

Equations (7a) and (7b) are, so far, exact, but they involve, on the right hand side, the functions E_t , H_t , E_z and H_z which are still unknown. We are interested in those cases where the integral terms are small, either as a consequence of the terms $(\epsilon_2 - \epsilon_1)$, η and so forth being small, or of their being finite only over a small region. In the first case the fields E_t , H_t , E_z and H_z may be replaced in the integrals by the values which they would have before the perturbation was made. In the second case this is not possible since a large change in the material constants of a

region alters the field substantially within that region. Here, then, we have a preliminary problem to solve, namely that of determining the field in the perturbed region in a zero order approximation.

Perturbation problems may be divided into two classes by another distinction. The changes in material properties may be independent of the z -coordinate, so that the new problem is to consider propagation in a uniform guide differing slightly from the original one. Typical of such problems is that of a waveguide containing a ferrite rod of infinite length parallel to the z -axis; the perturbation consisting here of the change in the properties of the rod when it is magnetized. Clearly, in such cases, solutions for which all field components vary as $\exp -j\beta z$ are still possible and the perturbation equations (7) become equations to determine β . On the other hand, there is a class of problems for which the perturbation is confined to a limited region in the z -direction, and we are interested, perhaps, in the reflection and transmission coefficients for a wave incident upon the obstacle. Here, for example, we might think of the case of a disc of ferrite across the guide. If we remain in the range for which perturbation theory is valid the changes in the amplitude of reflected and transmitted waves will be small, but the changes in phase may not be, if the perturbed region is sufficiently long. In the latter case, it would be possible, if the perturbation were uniform in z over the region in which it exists, to find solutions going as $\exp j\beta z$, as described above, and to use these to fit the boundary conditions at the ends. It is also possible if the perturbed region is long, with slowly varying properties, to obtain suitable approximate solutions by the WKB method. Some of these cases will arise in the examples which we treat below.

1.2. *Perturbations Uniform in z*

We consider first the general case in which the perturbation is uniform in z . In the absence of the perturbation the m^{th} mode is to be present. For the fields E_t and H_z , in the perturbed region we write $a_m(z)E_{tm0}(x, y)$ and $a_m(z)H_{zm0}(x, y)$, respectively, where $a_m(z)$ is the amplitude function for the m^{th} mode. If the perturbed region is one in which, for no magnetization, the material properties differ only slightly from their unperturbed values, we may justifiably identify E_{tm0} with E_{tm} and H_{zm0} with H_{zm} . If the material properties are appreciably changed even in the absence of a magnetic field, E_{tm0} and H_{zm0} have to be calculated by an independent method. For a_m we put $A_m e^{-j\beta z}$, where $\beta = \beta_m + \delta\beta$ and $\delta\beta$ is small. Similarly for H_t and E_z , we write $b_m H_{tm0}$ and $b_m E_{zm0}$, with $b_m = B_m e^{-j\beta z}$. With such assumptions, the m^{th} set of equations (7)

gives an equation for β , while any other set, with $n \neq m$, gives the excitation of the n^{th} mode. Substituting in (7) we have

$$j\beta_m A_m - j\beta B_m = \frac{j\omega}{\Delta_m} \left(\int [(\epsilon_2 - \epsilon_1) E_{tm0} \cdot \tilde{E}_{tm} - j\eta E_{tm0}^* \cdot \tilde{E}_{tm} + (\mu_3 - \mu_1) H_{zm0} \tilde{H}_{zm}] dS \right) A_m \equiv jL A_m \quad (8a)$$

and

$$j\beta_m B_m - j\beta A_m = \frac{j\omega}{\Delta_m} \left(\int (\mu_2 - \mu_1) H_{tm0} \cdot \tilde{H}_{tm} - j\kappa H_{tm0}^* \cdot \tilde{H}_{tm} + (\epsilon_3 - \epsilon_1) E_{zm0} \tilde{E}_{zm}] dS \right) B_m \equiv jM B_m. \quad (8b)$$

Ignoring squares and products of small quantities, one then has

$$\delta\beta = \frac{1}{2} (L + M). \quad (9)$$

The first example to be considered is the effect on the propagation in a circularly cylindrical waveguide, when a coaxial, magnetized pencil of ferrite of very small radius is introduced. The guide radius is r_0 and that of the pencil is r_1 . Before the ferrite is introduced, $\mu_1 = \mu_0$ and $\epsilon_1 = \epsilon_0$, where μ_0 and ϵ_0 are the free space values. The unperturbed fields are those of the usual TE and TM modes in round guide. It is necessary to calculate first the zero order electric and magnetic fields within the magnetized pencil; it will be sufficient to work out the magnetic case and deduce the electric one by analogy. Since the cross-section of the pencil is very small and transverse propagation effects consequently negligible, the internal field may be calculated by solving a static problem. The transverse magnetic field before the pencil is inserted is H_{tm} and it is assumed that the pencil is so small that over a circle with a few times the radius of the latter, H_{tm} is essentially uniform. We must now solve Laplace's equation for a cylindrical rod immersed in a magnetic field which is to be uniform at large distances. Within the rod, $B_t = \mu H_t - j\kappa H_t^*$, and at its surface the usual boundary conditions prevail. Hereafter we write μ for μ_2 .

The fields are derivable from potentials Φ_{out} , Φ_{in} , which are of the form

$$\Phi_{\text{in}} = (H_{tm0} \cdot \vec{r}),$$

$$\Phi_{\text{out}} = (H_{tm} \cdot \vec{r}) + \frac{(\vec{a} \cdot \vec{r})}{r^2},$$

where $\vec{r} \equiv (x, y)$, $\vec{a}$ is a constant vector and the coordinate system has its origin at the centre of the rod. Continuity of tangential H at the surface of the rod requires

$$H_{tm0} = H_{tm} + \frac{\vec{a}}{r_1^2}$$

and then

$$\Phi_{\text{out}} = H_{tm} \cdot \vec{r} + \frac{r_1^2}{r^2} (H_{tm0} - H_{tm}) \cdot \vec{r}.$$

The normal derivative at the surface of the rod is

$$\frac{1}{r_1} \left[x \frac{\partial \Phi}{\partial x} + y \frac{\partial \Phi}{\partial y} \right]$$

or, externally,

$$\frac{1}{r_1} [H_{tm} \cdot \vec{r} - (H_{tm0} - H_{tm}) \cdot \vec{r}].$$

Matching the normal B 's at the surface then gives

$$\mu_0 [2 H_{tm} - H_{tm0}] = \mu H_{tm0} - j\kappa H_{tm0}^*,$$

or

$$H_{tm0} = \frac{2\mu_0[(\mu + \mu_0)H_{tm} + j\kappa H_{tm}^*]}{(\mu + \mu_0)^2 - \kappa^2}. \quad (10a)$$

In a similar manner, one would find if the dielectric constant of the rod were ϵ ,

$$E_{tm0} = \frac{2\epsilon_0 E_{tm}}{\epsilon + \epsilon_0}. \quad (10b)$$

The longitudinal fields E_{zm} and H_{zm} are unchanged within the rod.

Turning now to the expression (9) for $\delta\beta$, we have in the present-case,

$$\begin{aligned} \delta\beta = & -\frac{\omega}{2\Delta_m} \int_{\text{pencil}} dS \left[\frac{2\epsilon_0(\epsilon - \epsilon_0)}{\epsilon + \epsilon_0} |E_{tm}|^2 + (\epsilon - \epsilon_0) |E_{zm}|^2 \right. \\ & \left. + 2\mu_0 \frac{\mu^2 - \mu_0^2 - \kappa^2}{(\mu + \mu_0)^2 - \kappa^2} |H_{tm}|^2 - j\kappa \frac{4\mu_0^2}{(\mu + \mu_0)^2 - \kappa^2} H_{tm}^* \tilde{H}_{tm} \right], \end{aligned}$$

where we have anticipated that Δ_m is real, which we verify below. Since the integrand is constant the integral may be replaced by πr_1^2 times the value of the integrand at $r = 0$. We consider now a TE-mode

with variation $e^{jn\varphi}$, $n \neq 0$. We shall have,

$$E_{tm} = -j\omega\mu_0 \nabla^* \Psi_m,$$

$$H_{tm} = -j\beta_m \nabla \Psi_m,$$

where

$$\Psi_m = J_n \left(u_{nm} \frac{r}{r_0} \right) e^{jn\varphi}, \quad J'_n(u_{nm}) = 0$$

and

$$\beta_m^2 = \omega^2 \epsilon_0 \mu_0 - \frac{u_{nm}^2}{r_0^2}.$$

For the fields on the axis, one finds for $n = \pm 1$,

$$nH_r = -jH_\varphi = -jn \frac{\beta_m u_{nm}}{2r_0} e^{jn\varphi}$$

and

$$nE_\phi = jE_r = j\omega_0 n \frac{u_{nm}}{2r_0} e^{jn\varphi}.$$

If $|n| \neq 1$ there is no first order perturbation. We now have

$$\delta\beta = -\frac{\omega}{2\Delta_m} \pi r_1^2 \left[\epsilon_0 \frac{\epsilon - \epsilon_0}{\epsilon + \epsilon_0} \omega^2 \mu_0^2 \frac{u_{nm}^2}{r_0^2} + \frac{\mu_0 \beta_m^2 u_{nm}^2}{r_0^2} \frac{\mu + n\kappa - \mu_0}{\mu + n\kappa + \mu_0} \right].$$

But we have

$$\begin{aligned} \Delta_m &= 2\pi \int_0^{r_0} E_{tm}^* \cdot \tilde{H}_{tm} r \, dr, \\ &= -2\pi \cdot \omega \mu_0 \beta_m \int_0^{u_{nm}} \left[J_n'(x)^2 + \frac{J_n(x)^2}{x^2} \right] x \, dx, \\ &= -2\pi \omega \mu_0 \beta_m \frac{J_n(u_{nm})^2}{2} (u_{nm}^2 - 1). \end{aligned}$$

and then,

$$\delta\beta = \frac{1}{2\beta_m} \cdot \frac{r_1^2}{r_0^2} \cdot \frac{u_{nm}^2}{J_n(u_{nm})^2 (u_{nm}^2 - 1)} \cdot \left[\beta_m^2 \frac{\mu + n\kappa - \mu_0}{\mu + n\kappa + \mu_0} + \beta_0^2 \frac{\epsilon_1 - \epsilon_0}{\epsilon_1 + \epsilon_0} \right]. \quad (11)$$

For TM modes we have

$$\begin{aligned} E_{tm} &= -j\beta_m \nabla \chi_m & E_z &= \frac{j_{nm}^2}{r_0^2} \chi_m \\ H_{tm} &= j\omega\epsilon_0 \nabla^* \chi_m \end{aligned}$$

where

$$\chi_m = J_n \left(j_{nm} \frac{r}{r_0} \right) e^{jn\varphi}, \quad J_n(j_{nm}) = 0$$

and

$$\beta_m^2 = \omega^2 \epsilon_0 \mu_0 - \frac{j_{nm}^2}{r_0^2}.$$

Proceeding as before, we find,

$$\delta\beta = \frac{1}{2\beta_m} \cdot \frac{r_1^2}{r_0^2} \cdot \frac{1}{J_n'(j_{nm})^2} \left[\beta_0^2 \frac{\mu}{\mu + n\kappa + \mu_0} + \beta_m^2 \frac{\epsilon - \epsilon_0}{\epsilon + \epsilon_0} \right]. \quad (12)$$

A problem which is of some interest, although not of immediate practical significance, is that of a ferrite pencil of arbitrary radius and infinite length in a round wave guide, with the remainder of the waveguide filled with a non-magnetic dielectric, whose dielectric constant, ϵ_1 , is equal to that of the ferrite. The ferrite is supposed to be only weakly magnetized. For such a problem, we have,

$$\delta\beta = -\frac{\omega}{2\Delta_m} \int_{\text{pencil}} [(\mu - \mu_0)H_{tm} \cdot \tilde{H}_{tm} - j\kappa H_{tm}^* \cdot \tilde{H}_{tm}] dS.$$

H_{tm} is the field of an unperturbed TE or TM mode in the dielectric-filled guide. $\mu - \mu_0$ and κ are supposed small, but r_1 , the radius of the pencil need no longer be small.

For TE modes, we have as before (again excluding the case $n = 0$),

$$\begin{aligned} E_{tm} &= -j\omega\mu_0 \nabla^* \Psi_m, \\ H_{tm} &= -j\beta_m \nabla \Psi_m, \\ \Psi_m &= J_n \left(u_{nm} \frac{r}{r_0} \right) e^{jn\varphi} \end{aligned}$$

and

$$\beta_m^2 = \omega^2 \epsilon_0 \mu_0 - \frac{u_{nm}^2}{r_0^2} = \beta_1^2 - \frac{u_{nm}^2}{r_0^2}.$$

Hence, assuming the pencil coaxial with the guide.

$$\begin{aligned}
 -\frac{2\Delta_m}{\omega} \delta\beta &= 2\pi \int_0^{r_1} (\mu - \mu_0) \beta_m^2 \left[\frac{u_{nm}^2}{r_0^2} J_n'^2 \left(u_{nm} \frac{r}{r_0} \right) \right. \\
 &\quad \left. + \frac{n^2}{r^2} J_n^2 \left(u_{nm} \frac{r}{r_0} \right) \right] r dr \\
 &\quad + 2\pi \kappa \beta_m^2 \cdot 2 \frac{u_{nm}}{r_0} n \int_0^{r_1} J_n \left(u_{nm} \frac{r}{r_0} \right) J_n' \left(u_{nm} \frac{r}{r_0} \right) dr, \\
 &= 2\pi \beta_m^2 \left[(\mu - \mu_0) \int_0^{u_{nm}(r_1/r_0)} \left[x J_n'^2(x) + \frac{J_n^2(x)}{x} \right] dx \right. \\
 &\quad \left. + 2\kappa n \int_0^{u_{nm}(r_1/r_0)} J_n(x) J_n'(x) dx \right], \\
 &= 2\pi \beta_m^2 \left[(\mu - \mu_0) \left(x J_n(x) J_n'(x) + \frac{x^2}{2} J_n'^2(x) \right. \right. \\
 &\quad \left. \left. + \frac{x^2 - 1}{2} J_n^2(x) \right) \right]_{x=u_{nm}(r_1/r_0)} + \kappa n J_n^2 \left(u_{nm} \frac{r_1}{r_0} \right) \Bigg].
 \end{aligned}$$

Making use of the value of Δ_m found in the preceding paragraphs, we have

$$\begin{aligned}
 \delta\beta &= \frac{\beta_m}{(u_{nm}^2 - 1) J_n(u_{nm})^2} \left[\left(\frac{\mu}{\mu_0} - 1 \right) \left(x J_n(x) J_n'(x) + \frac{x^2}{2} J_n'^2(x) \right. \right. \\
 &\quad \left. \left. + \frac{x^2 - 1}{2} J_n^2(x) \right) \right]_{x=u_{nm}(r_1/r_0)} + n \frac{\kappa}{\mu_0} J_n^2 \left(u_{nm} \frac{r_1}{r_0} \right) \Bigg]. \quad (13)
 \end{aligned}$$

For TM modes,

$$\begin{aligned}
 E_{tm} &= -j\beta_m \nabla \chi_m, \\
 H_{tm} &= j\omega \epsilon_1 \nabla^* \psi_m, \\
 \chi_m &= J_n \left(j_{nm} \frac{r}{r_0} \right) e^{jn\varphi}, \\
 \beta_m^2 &= \omega^2 \epsilon_1 \mu_0 - \frac{j_{nm}^2}{r_0^2} = \beta_1^2 - \frac{j_{nm}^2}{r_0^2}.
 \end{aligned}$$

In this case,

$$\begin{aligned}
 -\frac{2\Delta_m}{\omega} \delta\beta &= 2\pi (\omega \epsilon_1)^2 \left((\mu - \mu_0) \int_0^{r_1} \left[\frac{n^2}{r^2} J_n^2 \left(j_{nm} \frac{r}{r_0} \right) \right. \right. \\
 &\quad \left. \left. + \frac{j_{nm}^2}{r_0^2} J_n'^2 \left(j_{nm} \frac{r}{r_0} \right) \right] r dr + 2\kappa n \int_0^{r_1} \frac{j_{nm}}{r_0} J_n \left(j_{nm} \frac{r}{r_0} \right) J_n' \left(j_{nm} \frac{r}{r_0} \right) dr \right).
 \end{aligned}$$

The value of Δ_m for this case is

$$\Delta_m = -2\pi\omega\epsilon_1\beta_m \frac{j_{nm}^2}{2} J_n'^2(j_{nm}).$$

The value of $\delta\beta$ now becomes

$$\delta\beta = \frac{\beta_1^2}{\beta_m j_{nm}^2 J_n^2(j_{nm})} \left[\left(\frac{\mu}{\mu_0} - 1 \right) \left(x J_n(x) J_n'(x) + \frac{x^2}{2} J_n'^2(x) \right. \right. \\ \left. \left. + \frac{x^2 - 1}{2} J_n^2(x) \right)_{x=j_{nm}(r_1/r_0)} + n \frac{\kappa}{\mu_0} J_n^2 \left(j_{nm} \frac{r_1}{r_0} \right) \right]. \quad (14)$$

We note that for a ferrite filled guide with $r_1 = r_0$, the nonreciprocal part of $\delta\beta$ vanishes which confirms a result found in Part I of this paper for weak magnetization.

The very high dielectric constant of the ferrites (about 10) puts rather severe restrictions on the size of the pencils to which perturbation theory is applicable, even for weak magnetization. This limitation would be substantially relaxed if we possessed exact solutions for rods of high dielectric constant inserted into round guide, which could be used as the basis for magnetic perturbation calculations. Unfortunately, the only extensive published calculations of this kind are for dielectric constants less than 3. However, at the suggestion of M. T. Weiss, a calculation of the propagation constant of the lowest mode varying as $e^{\pm jz}$ in a wave guide containing a coaxial dielectric rod ($\epsilon_1 = 10$) has recently been made in the Mathematics Department, for varying rod diameter, but for a single value of guide radius equal to 0.4 times the free space wavelength. With the aid of this information, which was made available to us, the magnetic perturbation calculation has been carried out.

As before, the radius of the guide is r_0 and that of the rod is r_1 . The dielectric constant of the rod is ϵ_1 . We consider first the propagation in the unmagnetized case. Since we are considering only one mode, namely, that with an angular variation, $e^{j\phi}$, and of the lowest order radially, we need not identify the E 's and H 's by a label. We use a subscript "1" for fields in the dielectric and "0" for fields in the empty part of the guide. In general, we have

$$\alpha^2 E_t = -j[\omega\mu\nabla^* H_z + \beta\nabla E_z], \\ \alpha^2 H_t = -j[\beta\nabla H_z - \omega\epsilon\nabla^* E_z],$$

where

$$\alpha^2 = \omega^2\epsilon\mu - \beta^2, \\ \nabla^2 H_z = -\alpha^2 H_z, \\ \nabla^2 E_z = -\alpha^2 E_z.$$

and ϵ , μ refer, for the time being, to the dielectric constant and permeability of the region considered. It will be convenient to put

$$H_{z,t} = \sqrt{\frac{\epsilon_0}{\mu_0}} \bar{H}_{z,t}; \quad \frac{\epsilon}{\epsilon_0} = \bar{\epsilon}; \quad \frac{\mu}{\mu_0} = \bar{\mu};$$

$$\frac{\beta}{\omega \sqrt{\epsilon_0 \mu_0}} = \bar{\beta}; \quad \frac{\alpha}{\omega \sqrt{\epsilon_0 \mu_0}} = \bar{\alpha},$$

and to measure lengths in terms of $\frac{1}{\omega \sqrt{\mu_0 \epsilon_0}}$. We shall continue to use ∇ and ∇^* with the understanding that they refer to the scaled units.

We now have

$$\begin{aligned} j\bar{\alpha}^2 E_t &= \bar{\mu} \nabla^* \bar{H}_z + \bar{\beta} \nabla E_z, \\ j\bar{\alpha}^2 \bar{H}_t &= \bar{\beta} \nabla \bar{H}_z - \bar{\epsilon} \nabla^* E_z, \\ \bar{\alpha}^2 &= \bar{\epsilon} \bar{\mu} - \bar{\beta}^2, \\ \nabla^2 \bar{H}_z &= -\bar{\alpha}^2 \bar{H}_z, \\ \nabla^2 E_z &= -\bar{\alpha}^2 E_z. \end{aligned}$$

At the surface of the rod E_z , $\bar{H}_z$, E_φ and $\bar{H}_\varphi$ are continuous. We must have

$$\frac{1}{\bar{\alpha}_0^2} \left[\frac{j\bar{\beta}}{r_0} E_z(r_1) - \left(\frac{\partial \bar{H}_z}{\partial r} \right)_0 \right] = \frac{1}{\bar{\alpha}_1^2} \left[\frac{j\bar{\beta}}{r_1} E_z(r_1) - \left(\frac{\partial \bar{H}_z}{\partial r} \right)_1 \right]$$

and

$$\frac{1}{\bar{\alpha}_0^2} \left[\left(\frac{\partial E_z}{\partial r} \right)_0 + \frac{j\bar{\beta}}{r_1} \bar{H}_z(r_1) \right] = \frac{1}{\bar{\alpha}_1^2} \left[\bar{\epsilon}_1 \left(\frac{\partial E_z}{\partial r} \right)_1 + \frac{j\bar{\beta}}{r_1} \bar{H}_z(r_1) \right].$$

where

$$\bar{\epsilon}_1 = \frac{\epsilon_1}{\epsilon_0}, \quad \bar{\alpha}_0^2 = 1 - \bar{\beta}^2 \quad \text{and} \quad \bar{\alpha}_1^2 = \bar{\epsilon}_1 - \bar{\beta}^2.$$

$\bar{\mu} = 1$, everywhere, if the unmagnetized ferrite has the permeability of free space.

These equations may be rearranged in the form:

$$\begin{aligned} \bar{\beta} \left(\frac{1}{\bar{\alpha}_0^2} - \frac{1}{\bar{\alpha}_1^2} \right) E_z(r_1) &= \left\{ -\frac{1}{\bar{\alpha}_0^2} \left[r_1 \frac{1}{\bar{H}_z(r_1)} \left(\frac{\partial \bar{H}_z}{\partial r} \right)_0 \right] \right. \\ &\quad \left. + \frac{1}{\bar{\alpha}_1^2} \left[r_1 \frac{1}{\bar{H}_z(r_1)} \left(\frac{\partial \bar{H}_z}{\partial r} \right)_1 \right] \right\} j\bar{H}_z(r_1), \\ \bar{\beta} \left(\frac{1}{\bar{\alpha}_0^2} - \frac{1}{\bar{\alpha}_1^2} \right) j\bar{H}_z(r_1) &= \left\{ -\frac{1}{\bar{\alpha}_0^2} \left[r_1 \frac{1}{E_z(r_1)} \left(\frac{\partial E_z}{\partial r} \right)_0 \right] \right. \\ &\quad \left. + \frac{\bar{\epsilon}}{\bar{\alpha}_1^2} \left[r_1 \frac{1}{E_z(r_1)} \left(\frac{\partial E_z}{\partial r} \right)_1 \right] \right\} E_z(r_1). \end{aligned} \tag{15}$$

Within the dielectric, since all fields are bounded at $r = 0$, both E_z and $\bar{H}_z$ are proportional to $J_1(\bar{\alpha}_1 r)$ and, evidently,

$$-\frac{r_1}{H_z(r_1)} \left(\frac{\partial \bar{H}_z}{\partial r} \right)_1 = \frac{r_1}{E_z(r_1)} \left(\frac{\partial E_z}{\partial r_1} \right)_1 = \frac{\alpha_1 r_1 J_1'(\alpha_1 r_1)}{J_1(\alpha_1 r_1)} = F(\alpha_1 r_1),$$

in the notation of Part I. $(E_z)_0$ and $(H_z)_0$, similarly, will be those two linear combinations of $J_1(\bar{\alpha}_0 r)$ and $Y_1(\bar{\alpha}_0 r)$, which, respectively, vanish and have zero normal derivative at $r = r_0$, in order to ensure the vanishing of the tangential fields there. The functions

$$\frac{r}{(H_z)_0} \frac{\partial (H_z)_0}{\partial r} \quad \text{and} \quad \frac{r}{(E_z)_0} \frac{\partial (E_z)_0}{\partial r}$$

will be called $H(\bar{\alpha}_0 r)$ and $G(\bar{\alpha}_0 r)$ respectively.

Eliminating $E_z(r_1)$ and $H_z(r_1)$ from equations (15) we obtain the characteristic equation of the problem in the form

$$\bar{\beta}^2 \left(\frac{1}{\bar{\alpha}_0^2} - \frac{1}{\bar{\alpha}_1^2} \right)^2 = \left(\frac{F(\alpha_1 r_1)}{\bar{\alpha}_1^2} - \frac{H(\bar{\alpha}_0 r_1)}{\bar{\alpha}_0^2} \right) \left(\bar{\epsilon}_1 \frac{F(\bar{\alpha}_1 r_1)}{\bar{\alpha}_1^2} - \frac{G(\bar{\alpha}_0 r_1)}{\bar{\alpha}_0^2} \right).$$

The perturbation in the present problem is that due to a mild magnetization of the rod and referring again to equation (9) we have (in unscaled units)

$$\delta\beta = -\frac{\omega}{2\Delta_n} \int_{\text{rod}} [(\mu - \mu_0)H_t \cdot \bar{H}_t - j\kappa H_t^* \cdot \bar{H}_t] dS,$$

with

$$\Delta_n = \int_{\text{guide}} E_t^* \cdot \bar{H}_t dS.$$

Thus, in the scaled system,

$$\delta\bar{\beta} = -\frac{1}{2} \frac{\int_0^{r_1} \left[\left(\frac{\mu}{\mu_0} - 1 \right) |\bar{H}_t|^2 - j \frac{\kappa}{\mu_0} (\bar{H}_t^* \cdot \dot{\bar{H}}) \right] r dr}{\int_0^{r_0} E_t^* \cdot \dot{\bar{H}}_t r dr}. \quad (16)$$

The evaluation of the normalizing integral in the denominator is an exceedingly tedious business and it seems advisable to avoid it. This may be done in the following manner. The characteristic equation has been solved for numerous values of r_1 and $\bar{\beta}$ may be considered to be a reliably known function of r_1 . In particular the slope $\frac{d\bar{\beta}}{dr_1}$ is known. But we may also deduce $\frac{d\bar{\beta}}{dr_1}$, by a perturbation calculation in which we start with a

rod of radius r_1 and increase the latter to $r_1 + dr_1$ by changing ϵ_0 to ϵ_1 in the shell $r_1 < r < r_1 + dr_1$. For such a perturbation, since E_z and E_φ are continuous at the boundary, they suffer no change when the boundary moves; E_r , however is discontinuous and $(E_r)_o = \frac{\epsilon_1}{\epsilon_0}(E_r)_i$ where $(E_r)_i$ and $(E_r)_o$ are the normal fields just outside and just inside the rod. The perturbation formula (unscaled), thus gives

$$\delta\beta = \frac{-\omega}{2\Delta_n} (\epsilon_1 - \epsilon_0) \left(E_z^2 + E_\varphi^2 + \frac{\epsilon_1}{\epsilon_0} (E_r)_i^2 \right) \cdot 2\pi r_1 \delta r_1,$$

or, scaled, with r and r_1 representing scaled radii,

$$\delta\bar{\beta} = -\frac{1}{2} \frac{(\epsilon_1 - 1) (E_z^2 + E_\varphi^2 + \epsilon_1 (E_r)_i^2)}{\int_0^{r_0} E_t^* \cdot \tilde{H}_r r dr} \cdot r_1 \delta r_1.$$

The formula (16) for the magnetic perturbation may now be written

$$\delta\bar{\beta} = \frac{1}{\epsilon_1 - 1} \frac{\frac{d\bar{\beta}}{dr_1} \int_0^{r_1} \left[\left(\frac{\mu}{\mu_0} - 1 \right) |H_t|^2 - j \frac{\kappa}{\mu_0} (\tilde{H}_t^* \cdot \tilde{H}_t) \right] r dr}{E_z(r_1)^2 + |E_\varphi(r_1)|^2 + \epsilon_1 (E_r)_i^2}.$$

Inside the rod, we may write

$$\bar{H}_z = \frac{\bar{H}_z(r_1)}{E_z(r_1)} \quad E_z = jcE_z$$

and then

$$j\bar{\alpha}_1^2 \bar{H}_r = jc\bar{\beta} \frac{\partial E_z}{\partial r} - \epsilon_1 \frac{j}{r} E_z,$$

$$j\bar{\alpha}_1^2 \bar{H}_\varphi = jc\bar{\beta} \frac{j}{r} E_z + \epsilon_1 \frac{\partial E_z}{\partial r}.$$

The integrals are readily carried out and are as follows:

$$\begin{aligned} & \int_0^{r_1} |H_t|^2 r dr \\ &= \frac{E_z^2(r_1)}{\bar{\alpha}_1^4} \left[(\epsilon_1^2 + c^2 \bar{\beta}^2) \left(F(\bar{\alpha}_1 r_1) + \frac{F^2(\bar{\alpha}_1 r_1)}{2} + \frac{\bar{\alpha}_1^2 r_1^2 - 1}{2} \right) - 2c\bar{\beta}\epsilon_1 \right], \\ & \int_0^{r_1} (\tilde{H}_t^* \cdot \tilde{H}_t) r dr \\ &= \frac{jE_z^2(r_1)}{\bar{\alpha}_1^4} \left[(\epsilon_1^2 + c^2 \bar{\beta}^2) - 2c\bar{\epsilon}_1 \bar{\beta} \left(F(\bar{\alpha}_1 r_1) + \frac{F^2(\bar{\alpha}_1 r_1)}{2} + \frac{\bar{\alpha}_1^2 r_1^2 - 1}{2} \right) \right]. \end{aligned}$$

The term in the denominator may be evaluated by using

$$j\bar{\alpha}_1^2 E_r = -\frac{c}{r} E_z + \bar{\beta} \frac{\partial E_z}{\partial r},$$

$$j\bar{\alpha}_1^2 E_\phi = -jc \frac{\partial E_z}{\partial r} + \frac{j\bar{\beta}}{r} E_z.$$

We obtain

$$\begin{aligned} |E_z(r_1)|^2 + |E_\theta(r_1)|^2 + \bar{\epsilon}_1 |(E_r)_1|^2 \\ = E_z^2(r_1) \left[\frac{[\bar{\beta} - cF(\bar{\alpha}_1 r_1)]^2 + \bar{\epsilon}_1 [c - \bar{\beta}F(\bar{\alpha}_1 r_1)]^2}{r_1^2 \bar{\alpha}_1^4} \right]. \end{aligned}$$

The perturbation may now be written:

$$\begin{aligned} \delta\bar{\beta} = \frac{r_1 \frac{\partial \bar{\beta}}{\partial r_1}}{\bar{\epsilon}_1 - 1} \\ \cdot \frac{\left(\frac{\mu}{\mu_0} - 1 \right) [A(\bar{\epsilon}_1^2 + c^2 \bar{\beta}^2) - 2c\bar{\beta}\bar{\epsilon}_1] + \frac{\kappa}{\mu_0} [\bar{\epsilon}_1^2 + c^2 \bar{\beta}^2 - 2Ac\bar{\epsilon}_1\bar{\beta}]}{r_1^2 \bar{\alpha}_1^4 + [\bar{\beta} - cF(\bar{\alpha}_1 r_1)]^2 + \bar{\epsilon}_1 [c - \bar{\beta}F(\bar{\alpha}_1 r_1)]^2}, \end{aligned} \quad (17)$$

where

$$A = \frac{F^2(\bar{\alpha}_1 r_1)}{2} + F(\bar{\alpha}_1 r_1) + \frac{\bar{\alpha}_1^2 r_1^2}{2}$$

and c may be obtained from equation 15, and the definition of c

$$c = \frac{\frac{G(\bar{\alpha}_0 r_1)}{\bar{\alpha}_0^2} - \bar{\epsilon}_1 \frac{F(\bar{\alpha}_1 r_1)}{\bar{\alpha}_1^2}}{\bar{\beta} \left(\frac{1}{\bar{\alpha}_0^2} - \frac{1}{\bar{\alpha}_1^2} \right)} = \frac{\bar{\alpha}_1^2 G(\bar{\alpha}_0 r_1) - \bar{\epsilon}_1 \bar{\alpha}_0^2 F(\bar{\alpha}_1 r_1)}{\bar{\beta}(\bar{\epsilon}_1 - 1)}$$

with

$$G(\bar{\alpha}_0 r_1) = \bar{\alpha}_0 r_1 \frac{N_1(\bar{\alpha}_0 r_0) J_1'(\bar{\alpha}_0 r_1) - J_1(\bar{\alpha}_0 r_0) N_1'(\bar{\alpha}_0 r_1)}{N_1(\bar{\alpha}_0 r_0) J_1(\bar{\alpha}_0 r_1) - J_1(\bar{\alpha}_0 r_0) N_1(\bar{\alpha}_0 r_1)}.$$

Fig. 1 shows the propagation constant as a function of the relative diameter, r_1/r_0 , of the dielectric rod in the unmagnetized case, with $\bar{\epsilon}_1 = 10$. Fig. 2 shows the derivative $\frac{d\bar{\beta}}{d(r_1/r_0)}$ as a function of r_1/r_0 . The guide radius is 0.4 times the free space wavelength. From equation (17) $\delta\bar{\beta}$ may be written as

$$\delta\bar{\beta} = P(\mu - \mu_0) + Q\kappa.$$

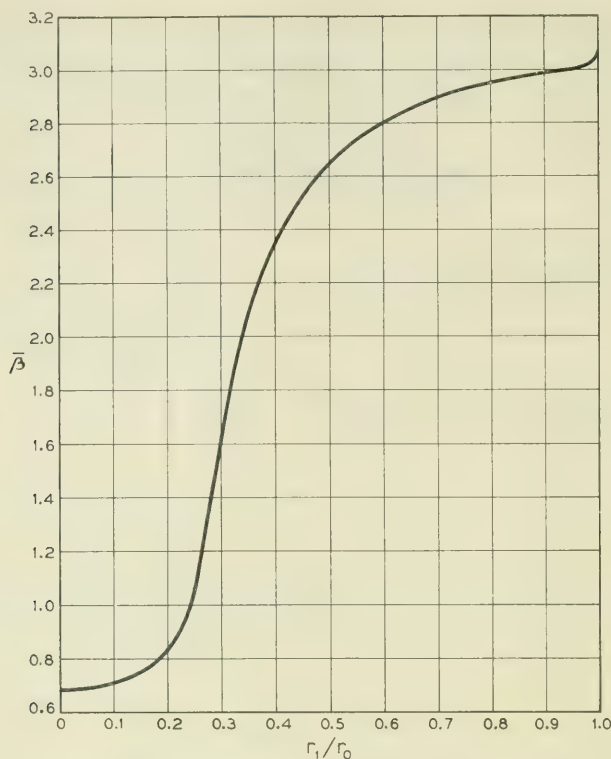


Fig. 1 — Propagation constant of a circular guide containing an unmagnetized coaxial, dielectric rod ($\epsilon/\epsilon_0 = 10$). r_1 is the radius of the rod and r_0 the radius of the guide. $r_0 = 0.4\lambda_0$, where λ_0 is the free space wavelength.

The computed values of Q and $P - Q$ are shown in Fig. 3 as a function of relative rod radius. $P - Q$ is plotted since P and Q are very nearly equal.*

1.3 Perturbations Non-Uniform in z

So far we have been concerned only with structures indefinitely extended in the direction of wave propagation. In practice the non-reciprocal element is, of course, finite, but end effects can frequently be neglected, since the element is matched at its ends (by tapering of the finite

* H. Seidel and Miss M. J. Brannon, at the suggestion of M. T. Weiss, have recently calculated the dielectric loss for the guides containing a dielectric rod described above. By combining such information with that obtained here it is possible to discuss figures of merit (degrees of rotation loss in db) for various pencil radii. Such an analysis is being made by M. T. Weiss and will appear in an article, by S. E. Miller, A. G. Fox and M. T. Weiss, in a forthcoming issue of the JOURNAL.

sample, for instance). The matching could be accomplished in such a way that the transition region, whose characteristics would be very difficult to compute, should contribute little to the overall non-reciprocal behavior. Therefore, in many cases, the theory for the indefinitely extended sample is adequate. For some special purposes, however, it is desirable to mismatch the sample deliberately. For instance, Rowen³ has suggested that the change in Faraday rotation, due to internal reflections in an unmatched specimen, can offset to some extent the frequency dependence of the rotation which is implied by the Polder relations, broadening thereby the useful bandwidth of the device.

Consider an infinite slab of ferrite, magnetized in a direction normal to its two parallel plane bounding surfaces. A circularly polarized wave, normally incident on the slab, will be partially transmitted, and, since for such a wave, the medium behaves as though it had an ordinary scalar permeability, the phase and amplitude of the transmitted portion are readily calculated. It is clear that, as the result of multiple internal reflections, the phase of the emerging wave will differ from the value appropriate to a single trip through the slab (such as would be obtained were the slab perfectly matched). Both amplitude and phase of the transmitted wave will depend on the electrical thickness of the slab and its dielectric constant, and on the effective permeability. The latter

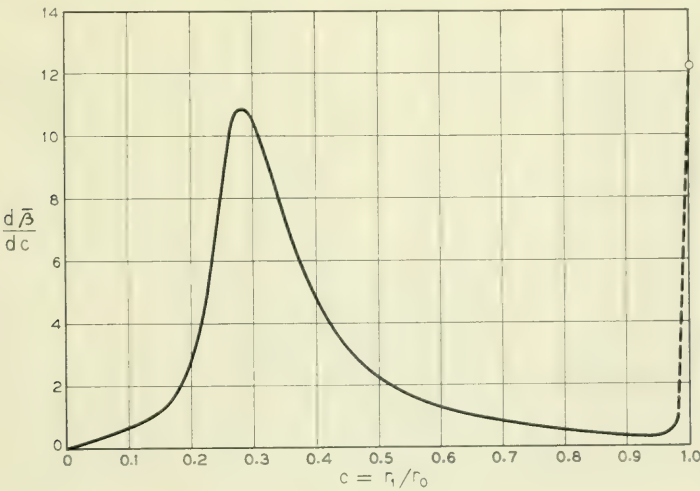
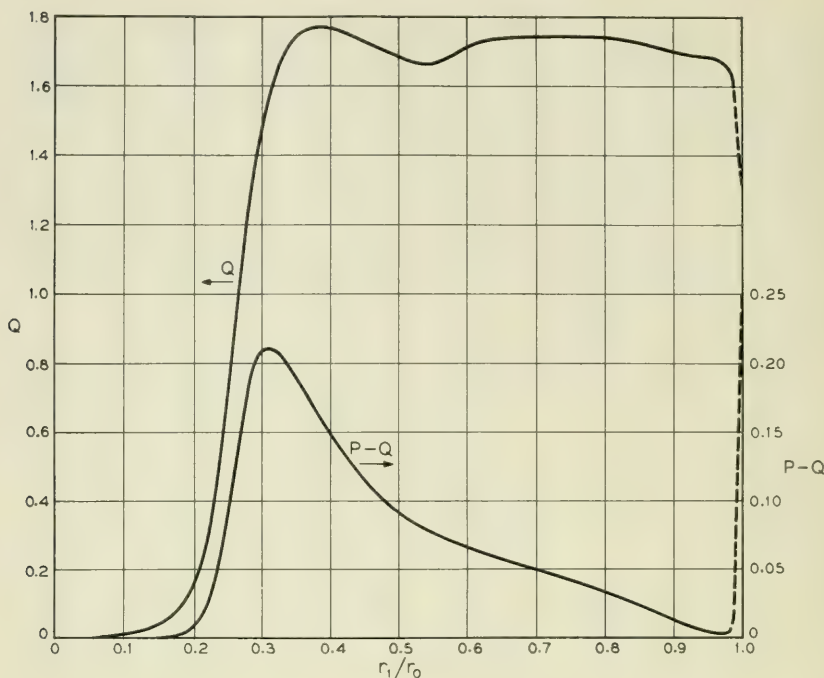


FIG. 2 — $\frac{d\beta}{d\left(\frac{r_1}{r_0}\right)}$ versus $\frac{r_1}{r_0}$.

Fig. 3 — Q and $P - Q$ versus r_1/r_0 .

differs for right and left circular waves, so that a plane polarized signal (which is the sum of equal right and left circular components) will emerge, in general, elliptically polarized, with the major axis of the ellipse tilted from the polarization at incidence by an angle which differs from the single trip value as the result of internal reflections. It is clear that this change in rotation can be calculated in a very elementary way.

When the sample is confined to a waveguide a similar effect occurs, but its calculation becomes extremely involved, at least for arbitrarily large magnetizations. The reason, which should be clear from Part I of this paper, is that the circularly polarized modes no longer have the same field configurations inside and outside the sample.* Therefore any incident mode excites *all* of the normal modes of the ferrite; these, in turn, give rise to *all* the mode patterns of the air-filled portion of the guide. Even if all but one of these are cut off, the excitation modifies the phase and amplitude of the reflected and transmitted portions of the propagating mode.

* This is due to the fact that there is now no ordinary effective scalar permeability as for infinite geometry.

Thus all modes have to be included in the problem, which consequently takes the form of an infinite system of linear equations for the mode amplitudes. This can be solved only to some approximation whose general validity it would be hard to establish. The problem could also be stated as an integral equation involving a complicated Green's function, with no greater chance of complete solution.

We are therefore forced to restrict the problem to the ranges of magnetization, or of sample size, in which perturbation theory is applicable. However, we begin with a discussion of the infinite, plane, loss-free slab, a problem which can be solved completely, and which has some bearing on the perturbation problem for a slug of ferrite whose cross section completely fills the waveguide.

Let the plane boundaries of the slab be normal to the z -axis, which is also the direction of magnetization and the propagation direction of a circularly polarized wave incident on the boundary $z = 0$ (see Fig. 4). In terms of the parameters p and σ of Section 2.1, Part I, the effective permeability for a circularly polarized wave is

$$\mu \pm \kappa = \mu = \mu_0 \left(1 \pm \frac{p}{1 \pm \sigma} \right),$$

the upper sign referring to right circular polarization. The corresponding propagation constants in the slab are then

$$\beta_{\pm} = \omega \sqrt{\epsilon \mu_0} \sqrt{1 \pm \frac{p}{1 \pm \sigma}}$$

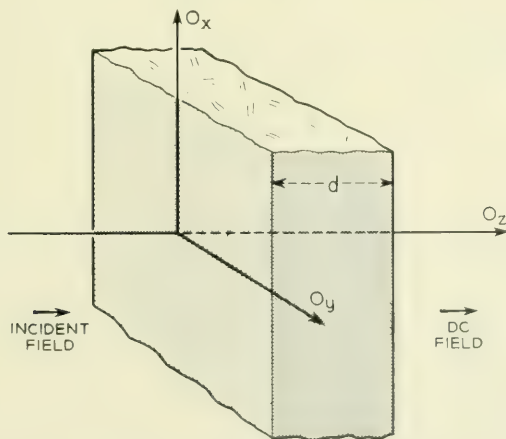


Fig. 4 — Normally magnetized ferrite slab.

where ϵ is the dielectric constant of the ferrite. If d is the thickness of the slab, the electrical thickness is

$$\theta_{\pm} = \beta_{\pm} d = \frac{2\pi d}{\lambda_{\pm}} = \theta_0 \sqrt{1 \pm \frac{p}{1 \pm \sigma}},$$

where $\theta_0 = \omega \sqrt{\mu_0 \epsilon} d$ is the electrical thickness of the unmagnetized sample. Let us now confine the discussion to the right circular wave. If the incident electric field is taken to be $e^{-j\beta_0 z}$, where β_0 is the free space propagation constant, $\omega \sqrt{\mu_0 \epsilon_0}$, the incident magnetic field will be

$$-\frac{\beta_0}{\omega \mu_0} e^{-j\beta_0 z},$$

and if the reflected electric field is $\rho e^{j\beta_0 z}$, the reflected magnetic field will be

$$\frac{\beta_0}{\omega \mu_0} \rho e^{j\beta_0 z}$$

since β_0 reverses sign. Inside the slab, the electric field consists of forward and backward travelling parts $\tau_1 e^{-j\beta_+ z}$ and $\tau_2 e^{j\beta_+ z}$, and the corresponding magnetic fields are

$$-\frac{\beta_+}{\omega \mu_+} \tau_1 e^{-j\beta_+ z}$$

and

$$\frac{\beta_+}{\omega \mu_+} \tau_2 e^{j\beta_+ z}.$$

Finally the transmitted electric and magnetic fields will be denoted by

$$\tau_3 e^{-j\beta_0(z-d)}$$

and

$$-\frac{\beta_0}{\omega \mu_0} \tau_3 e^{-j\beta_0(z-d)}$$

respectively. In general ρ , as well as the τ 's, will be complex. To obtain τ_3 , we write down the equations of continuity of all tangential fields across the boundaries $z = 0$ and $z = d$. Since the fields are confined to a plane normal to the z -direction, these equations are:

$$1 + \rho = \tau_1 + \tau_2,$$

$$\frac{\beta_0}{\omega \mu_0} (1 - \rho) = \frac{\beta_+}{\omega \mu_+} (\tau_1 - \tau_2)$$

and

$$\begin{aligned}\tau_3 &= \tau_1 e^{-j\beta_+ d} + \tau_2 e^{-j\beta_+ d}, \\ \frac{\beta_0}{\omega \mu_0} \tau_3 &= \frac{\beta_+}{\omega \mu_0} (\tau_1 e^{-\beta_+ d} - \tau_2 e^{j\beta_+ d}).\end{aligned}$$

These four equations in four unknowns are easily solved for τ_3 . Writing

$$x_+ = \sqrt{1 + \frac{p}{1 + \sigma}}$$

and noting that

$$\begin{aligned}\frac{\beta_+}{\mu_+} \bigg/ \frac{\beta_0}{\mu_0} &= \sqrt{\frac{\epsilon}{\epsilon_0}} \bigg/ x_+, \\ &= \frac{a}{x_+}, \quad \text{say,}\end{aligned}$$

where

$$a = \sqrt{\frac{\epsilon}{\epsilon_0}},$$

one finds the solution to be

$$\begin{aligned}\tau_3 &= \frac{1}{\cos \theta_0 x + \frac{j}{2} \left(\frac{a}{x_+} + \frac{x_+}{a} \right) \sin \theta_0 x_+}, \\ &= |\tau_3|_+ e^{-j\Phi_+},\end{aligned}$$

where

$$\tan \Phi_+ = \frac{1}{2} \left(\frac{a}{x_+} + \frac{x_+}{a} \right) \tan \theta_0 x_+ \quad (18)$$

and

$$|\tau_3|_+ = \left[\cos^2 \theta_0 x_+ + \frac{1}{4} \left(\frac{a}{x_+} + \frac{x_+}{a} \right)^2 \sin^2 \theta_0 x_+ \right]^{-1/2}. \quad (19)$$

Similar results apply to a left circular wave; it is necessary only to reverse the signs of σ , p in the expression for x_+ . Equations (18) and (19) show that equal right and left circular incident waves emerge with different amplitudes and phases; hence an incident plane polarized wave emerges elliptically polarized with its major axis inclined to the polarization direction at incidence. The inclination and the ratio of minor to major axis

are determined as follows: The right and left circular fields, upon emergence, may be written in terms of rectangular components (with the polarization of the total incident field along the x -axis)

$$E_x^+ - jE_y^+ = \tau_+ e^{j\omega t},$$

$$E_x^- + jE_y^- = \tau_- e^{j\omega t},$$

from which the resultant field in the x -direction is seen to be

$$\left. \begin{aligned} E_{xT} &= E_x^+ + E_x^- = |\tau_+| \cos(\omega t - \Phi_+) + |\tau_-| \cos(\omega t - \Phi_-) \\ \text{and in the } y \text{ direction} \\ E_{yT} &= E_y^+ + E_y^- = |\tau_-| \sin(\omega t - \Phi_-) - |\tau_+| \sin(\omega t - \Phi_+). \end{aligned} \right\} \quad (20)$$

The amplitude at time t , $(E_{xT}^2 + E_{yT}^2)^{1/2}$, is thus given by

$$\begin{aligned} E_{xT}^2 + E_{yT}^2 &= |\tau_+|^2 + |\tau_-|^2 \\ &\quad + 2|\tau_+| \cdot |\tau_-| \cos[2\omega t - (\Phi_- + \Phi_+)] \quad (21) \end{aligned}$$

The major axis of the ellipse is the maximum of $(E_{xT}^2 + E_{yT}^2)^{1/2}$ with respect to ωt . It equals $|\tau_+| + |\tau_-|$ and is attained at

$$\omega t = \frac{1}{2}(\Phi_- + \Phi_+).$$

Similarly the minor axis is the minimum and equals $|\tau_+| - |\tau_-|$. The ratio of minor to major axis is therefore

$$\frac{|\tau_+| - |\tau_-|}{|\tau_+| + |\tau_-|}.$$

The angle between the x -axis (the incident polarization direction) and the major axis is found by substituting $\omega t = \frac{1}{2}(\Phi_- + \Phi_+)$ in (20). This gives

$$E_{xT} = (|\tau_+| + |\tau_-|) \cos \frac{\Phi_+ - \Phi_-}{2},$$

$$E_{yT} = (|\tau_+| + |\tau_-|) \sin \frac{\Phi_+ - \Phi_-}{2},$$

which shows that the angle is

$$\frac{\Phi_+ - \Phi_-}{2}.$$

$|\tau|$ and Φ are plotted versus x in Fig. 5. τ_+ , Φ_+ and τ_- , Φ_- at given

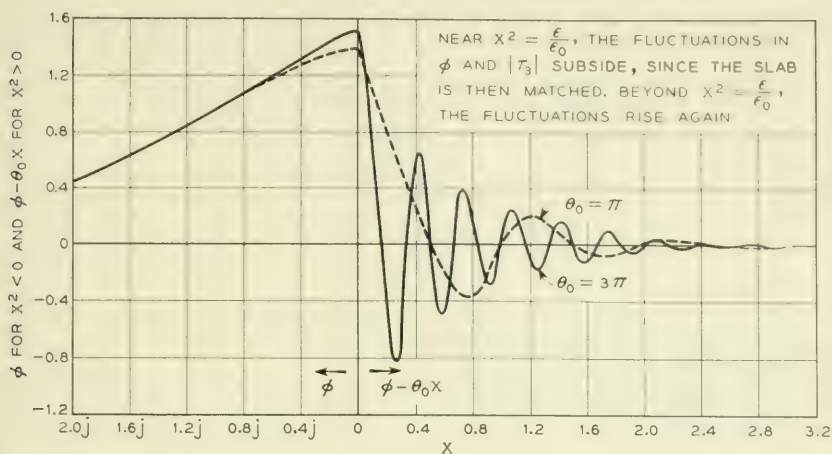


FIG. 5a

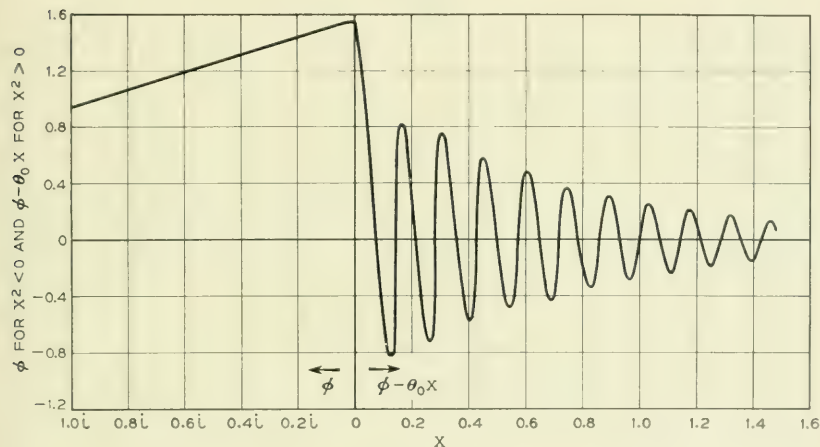


FIG. 5b

Fig. 5 — Phase and amplitude of transmitted circularly polarized wave as a function of $x = \sqrt{1 + \frac{p}{1 + \sigma}}$, with $\frac{\epsilon}{\epsilon_0} = 10$. (a), top, Φ for $\theta_0 = \pi$ and 3π ; (b), bottom, Φ for $\theta_0 = 7\pi$; (c), $|\tau_3|$ for $\theta_0 = \pi$ and 3π ; and (d), $|\tau_3|$ for $\theta_0 = 7\pi$.

$|\sigma|$, $|p|$ are found by choosing positive and negative signs for σ and p in

$$x = \sqrt{1 + \frac{p}{1 + \sigma}}.$$

Note that x can be imaginary corresponding to cut-off in the range $-1 < \sigma < -1 - p$ ($p < 0$).

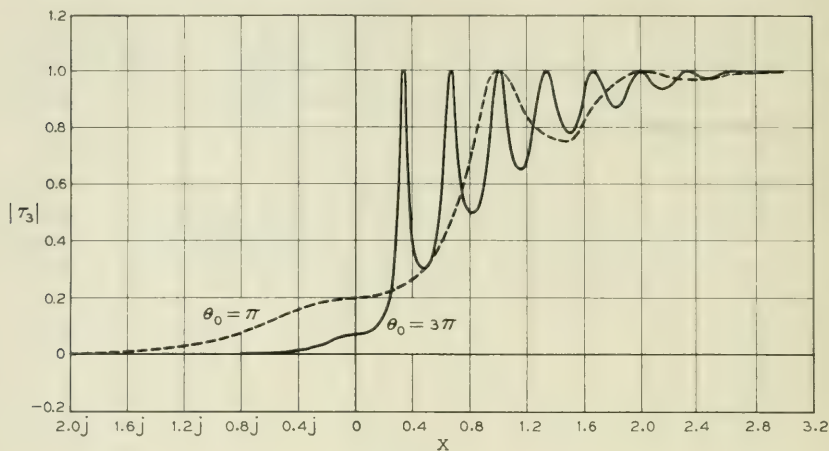


FIG. 5c

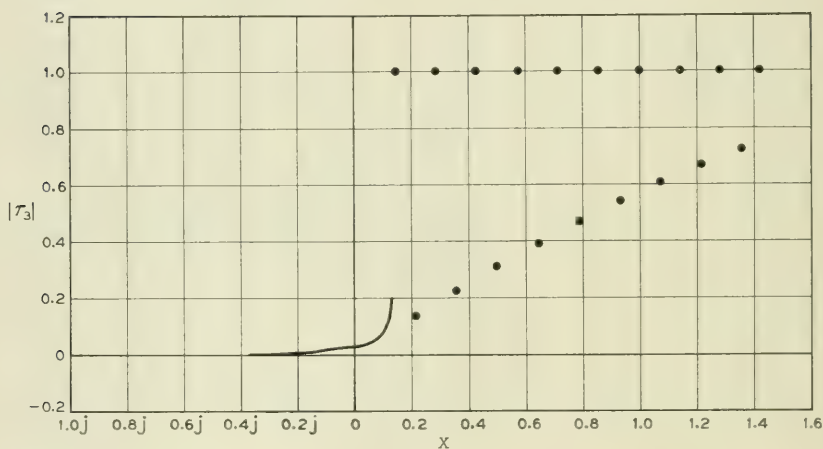


FIG. 5d

Fig. 5(c), top and (d), bottom — See Fig. 5.

Near $x = 1$ (corresponding either to small p or to sufficiently large σ), Φ differs from its value Φ_0 for the isotropic case by a small amount $\delta\Phi$. The rotation, to first order, is then just one half the difference between the $\delta\Phi$ for positive and negative p , σ . Writing

$$x_{\pm} = 1 + \delta x_{\pm} = 1 \pm \frac{1}{2} \frac{p}{1 \pm \sigma}$$

and expanding equation (18), we obtain

$$\delta\Phi_{\pm} = \theta_0 \delta x_{\pm} \frac{\cosh \Psi \sec^2 \theta_0 - \frac{\tan \theta_0}{\theta_0} \sinh \Psi}{1 + \cosh^2 \Psi \tan^2 \theta_0},$$

where Ψ is defined by

$$a = \sqrt{\frac{\epsilon}{\epsilon_0}} = e^{\Psi}.$$

[This result holds even near $\theta_0 = (n + \frac{1}{2})\pi$ where $\tan \theta_0 = \infty$, as can be seen by expanding the reciprocal of equation (18).] The quantity $\frac{1}{2}\theta_0(\delta x_+ - \delta x_-)$ is the rotation corresponding to a single trip through the sample. The actual rotation is $\frac{1}{2}(\delta\Phi_+ - \delta\Phi_-)$. Hence we may define a *rotation gain* as the ratio

$$\begin{aligned} g(\theta_0, a) &= \frac{\frac{1}{2}(\delta\Phi_+ - \delta\Phi_-)}{\frac{1}{2}\theta_0(\delta x_+ - \delta x_-)}, \\ &= \frac{\cosh \Psi \sec^2 \theta_0 - \frac{\tan \theta_0}{\theta_0} \sinh \Psi}{1 + \cosh^2 \Psi \tan^2 \theta_0}. \end{aligned} \tag{22}$$

In many cases, $\theta_0 \gg 1$, that is, the thickness of the sample is much greater than a reduced wavelength in the specimen. Then the second term in the numerator of g is always negligible compared with the first. g then simplifies to

$$g_1 = \frac{\cosh \Psi}{\cos^2 \theta_0 + \cosh^2 \Psi \sin^2 \theta_0}.$$

This expression is plotted in Fig. 6 as a function of θ for various $a = e^{\Psi}$. For given Ψ it has minima equal to $\frac{1}{\cosh \Psi}$ at $\theta_0 = \left(n + \frac{1}{2}\right)\pi$, and maxima equal to $\cosh \Psi$ at $\theta_0 = n\pi$ ($n = 0, 1, 2, \dots$). When $a \gg 1$ ($a \sim 3$ for many ferrites), $\cosh \Psi$ is replaced by $\frac{1}{2}e^{\Psi} = \frac{1}{2}a$, and then

$$\begin{aligned} g_{1\max} &= \frac{1}{2}a \\ g_{1\min} &= \frac{2}{a} \end{aligned}$$

It is to be noted that when $\theta = n\pi$, the condition for maximum g_1 , the unperturbed reflected amplitude is zero, and the ellipticity vanishes to first order in δx .

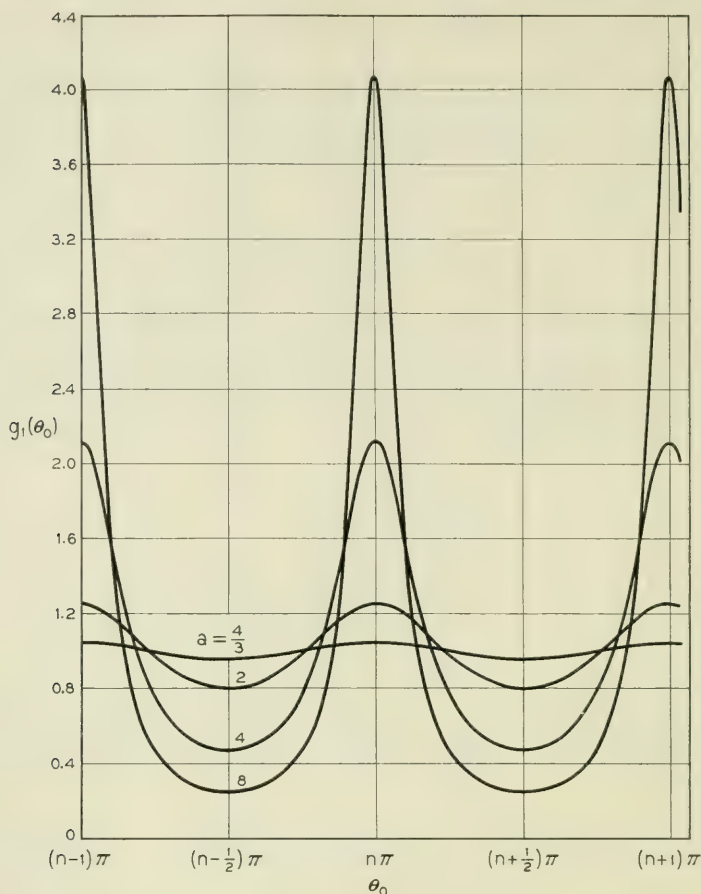


Fig. 6 — Rotation gain, $g_1(\theta_0)$, versus electrical thickness, θ_0 , for small magnetization.

Perturbation theory enables us to solve approximately, for weak magnetization, the problem just considered for the case in which a right circular cylinder of ferrite snugly fits a cylindrical waveguide. We will show that, with a suitable reinterpretation of the constants, equation (22) for the rotation gain of the extended slab continues to apply here.

Before magnetization, a particular right circularly polarized mode, say the m^{th} , is present in the sample. The small magnetization distorts the pattern of this mode and changes the propagation constant slightly. The distorted pattern can be expanded in the series of normal modes of the unperturbed material. In these expansions only the coefficient of the

originally present m^{th} mode will be large; all others will at most be of the order of the perturbation. Denoting perturbed quantities by the superfix +, we have for the distorted fields

$$\begin{aligned} E_{tm}^+ &= e^{-j\beta_{tm}^+ z} \sum_{n=1}^{\infty} p_{mn} E_{tn}, \\ H_{tm}^+ &= e^{-j\beta_{tm}^+ z} \sum_{n=1}^{\infty} q_{mn} H_{tn}, \end{aligned}$$

where p_{mm} and q_{mm} are large compared with the other coefficients. β_m^+ and the p 's and q 's are determined from equations (7a) and (7b). a_n in these equations is identified with $e^{-j\beta_{mn}^+ z} p_{mn}$, b_n with $e^{-j\beta_{mn}^+ z} q_{mn}$. Since all perturbation integrals involving E_t and H_z , E_z vanish in the present case, we obtain

$$\beta_n q_{mn} - \beta_m^+ p_{mn} = \frac{\omega}{\Delta_n} \left[\int (\mu - \mu_0) H_{tm}^+ \tilde{H}_{tn} dS - j \int \kappa H_{tm}^{+*} \tilde{H}_{tn} dS \right]$$

and

$$\beta_n p_{mn} - \beta_m^+ q_{mn} = 0.$$

In the first of these equations, H_{tm}^+ , the perturbed magnetic field, consists of $q_{mm} H_{tm}$, plus an admixture of other modes with coefficients themselves of order $\mu - \mu_0$, κ . Therefore, to first order, it suffices to write

$$H_{tm}^+ = q_{mm} H_{tm}$$

in the integrand, with the result:

$$\begin{aligned} \beta_n q_{mn} - \beta_m^+ p_{mn} &= I_{mn}^+ q_{mm}, \\ p_{mn} &= \frac{\beta_m^+}{\beta_n} q_{mn}, \end{aligned}$$

where

$$I_{mn}^+ = \frac{\omega}{\Delta_n} \int (\mu - \mu_0) H_{tm} \tilde{H}_{tn} - j \kappa H_{tm}^* \tilde{H}_{tn} dS.$$

Elimination of p_{mn} gives

$$(\beta_n^2 - \beta_m^{+2}) q_{mn} = \beta_n I_{mn}^+ q_{mm}.$$

The case $n = m$ determines β_m^{+2} :

$$\beta_m^{+2} = \beta_m^2 (1 - I_{mm}^+ \beta_m). \quad (23)$$

All other cases give, to first order,

$$q_{nm} = \frac{\beta_n I_{mn}^+}{\beta_n^2 - \beta_m^2} q_{mm},$$

which, again to first order, equals

$$\frac{\beta_n I_{mn}^+}{\beta_n^2 - \beta_m^2} p_{mm}. \quad (24)$$

One of the quantities p_{mm} , q_{mm} , is still arbitrary. If it is required that the perturbed field is, to zero order in I_{mn}^+ , normalized to the same value as the unperturbed field, it is readily found that one can take

$$p_{mm} = 1, \quad q_{mm} = \frac{\beta_m}{\beta_m^+}.$$

We are now in a position to consider the problem of a right circular mode, say the r^{th} , incident upon the end plane, $z = 0$, of the ferrite cylinder extending to $z = d$. One simplifying feature of this problem is that the unperturbed modes inside, and the modes outside, the sample have the same dependence on radius since the sample fills the whole guide-cross-section. However the modes inside and outside may have different numerical coefficients. Thus, if we distinguish quantities outside the sample by primes, the TE_{sr} mode can be represented outside and inside the sample by

$$\begin{aligned} E'_{t_{sr}} &= E_{t_{sr}} = \nabla^* \psi_{sr}, \\ H'_{t_{sr}} &= \frac{\beta'_r \operatorname{sgn} \beta'_r}{\omega \mu_0} \nabla \psi_{sr}, \\ H_{t_{sr}} &= \frac{\beta_r \operatorname{sgn} \beta_r}{\omega \mu_0} \nabla \psi_{sr} \end{aligned}$$

Here $\operatorname{sgn} \beta$ means: sign of the propagation direction, $+1$ and -1 for forward and reverse respectively. The function ψ_{sr} is given by

$$\psi_{sr} = J_s \left(u_r \frac{r}{r_0} \right) e^{js\varphi},$$

where φ is the azimuthal angle, r_0 the guide radius, and u_r the r^{th} zero of $J'_s(x)$. For TM modes we have similarly

$$\begin{aligned} E'_{t_{sr}} &= E_{t_{sr}} = \nabla \psi_{sr}, \\ H_{t_{sr}} &= -\frac{\omega \epsilon_0}{\beta_r} (\Delta^* \psi_{sr}) \operatorname{sgn} \beta_r, \\ H_{t_{sr}} &= -\frac{\omega \epsilon_1}{\beta_r} (\nabla^* \psi_{sr}) \operatorname{sgn} \beta_r, \end{aligned}$$

where in the definition of the ψ , u_r is replaced by j_r , the r^{th} zero of $J_s(x)$. Since the perturbations considered here do not couple modes with different azimuthal number, s will be considered fixed hereafter, and only the suffix r will be retained.

The field at $z = -0$ (just outside the ferrite) will consist, not only of the incident field and its reflection ρ_r , but also of other modes excited by the perturbation. Taking the incident E'_r to have unit amplitude, we have, from the continuity of tangential electric fields,

$$\begin{aligned} (1 + \rho_r)E'_{tr} + \sum_{n \neq r} \rho_n E'_{tn} &= \sum_n (\tau_{n1} + \tau_{n2}) \sum_{\ell} p_{n\ell} E_{\ell t}, \\ &= \sum_{n, \ell} (\tau_{n1} + \tau_{n2}) p_{n\ell} E_{\ell t}, \end{aligned} \quad (25)$$

where the τ_{n1} , τ_{n2} are respectively the forward and backward traveling amplitudes of the n^{th} perturbed mode. In the same way, we have for the tangential magnetic field:

$$(1 - \rho_r)H'_{tr} - \sum_{n \neq r} \rho_n H'_{tn} = \sum_{n, \ell} (\tau_{n1} - \tau_{n2}) q_{n\ell} H_{\ell t}. \quad (26)$$

The changes in sign of the coefficients of the backward waves have already been explained in Section 1.1. Let us now suppose that the incident mode is the TE_r mode. Multiplying both sides of (25) by $\nabla^* \tilde{\psi}_r$ and both sides of (26) by $\nabla \tilde{\psi}_r$, and integrating over the cross-section we have, from the orthogonality relations of these functions,

$$\begin{aligned} 1 + \rho_r &= \sum_n (\tau_{n1} + \tau_{n2}) p_{nr}, \\ \beta'_r (1 - \rho_r) &= \sum_n \beta_r (\tau_{n1} - \tau_{n2}) q_{nr}. \end{aligned}$$

But in these series, all τ except τ_r will be small, as will be all p , q except p_{rr} , q_{rr} . Hence all terms, except those for which $n = r$, will be small of second order, and can be neglected. Further,

$$p_{rr} = 1, \quad q_{rr} = \frac{\beta_r}{\beta_r^+}.$$

Thus we obtain finally,

$$\begin{aligned} 1 + \rho_r &= \tau_{r1} + \tau_{r2}, \\ 1 - \rho_r &= (\tau_{r1} - \tau_{r2}) \frac{\beta_r^2}{\beta'_r \beta_r^+}. \end{aligned}$$

In view of equation (23) for β_r^+ , these equations can be written, to

first order in I_{rr}^+ ,

$$\begin{aligned} 1 + \rho_r &= \tau_{r1} + \tau_{r2} \\ 1 - \rho_r &= \frac{\beta_r^+ \mu_0}{\beta_r' \mu_{r+}} (\tau_{r1} - \tau_{r2}) \end{aligned} \quad (27)$$

where

$$\mu_{r+} = \mu_0 \left(1 - \frac{I_{rr}^+}{\beta_r} \right).$$

Similarly at the output plane, $z = d$, the transmitted amplitude τ_{r3} satisfies

$$\begin{aligned} \tau_{r3} &= \tau_{r1} e^{-j\theta} + \tau_{r2} e^{j\theta}, \\ \tau_{r3} &= \frac{\beta_r^+ \mu_0}{\beta_r' \mu_{r+}} (\tau_{r1} e^{-j\theta} - \tau_{r2} e^{j\theta}), \end{aligned} \quad (28)$$

where $\theta = \beta_r^+ d$. Comparison of equations (27) and (28) with the corresponding equations for the infinite slab shows that they have the same form; however, equations (27) and (28) hold only for small magnetization. Hence only those results for the slab that relate to small magnetization can be generalized to the present case. One such result is that for the rotation gain. If we re-identify variables by

$$\begin{aligned} \delta x^+ &\rightarrow \delta x_r^+ = \frac{1}{2} \frac{I_{rr}^+}{\beta_r} = -\frac{1}{2} \frac{I_{rr}^-}{\beta_r}, \\ \theta_0 &\rightarrow \theta_r = \beta_r d, \\ a &\rightarrow a_r = \beta_r / \beta_r', \end{aligned}$$

we obtain, for the rotation gain,

$$g_r = \frac{\cosh \psi_r \sec^2 \theta_r - \frac{\tan \theta_r}{\theta_r} \sinh \psi_r}{1 + \cosh^2 \psi_r \tan^2 \theta_r}, \quad (29)$$

where

$$e^{\psi_r} = a_r = \beta_r / \beta_r'.$$

The formulae and conditions for maximum and minimum rotation gain derived in the previous section, apply here also, in terms of the re-defined variables.

The conversion of part of the incident mode to the s^{th} mode in the transmitted wave can be examined by multiplying the matching equa-

tions by $\nabla^* \tilde{\psi}_s$ or $\nabla \tilde{\psi}_s$ and integrating over the cross-section. It is then found that the transmitted amplitude of the n^{th} mode is

$$\tau_{n3} = 2\gamma_{rn} Z_n \frac{(Z_r + Z_n)(\cos \theta_n - \cos \theta_r) + j(1 + Z_n Z_r)(\sin \theta_n - \sin \theta_r)}{(2Z_n \cos \theta_n + j(Z_n^2 + 1) \sin \theta_n)(2Z_r \cos \theta_r + j(Z_r^2 + 1) \sin \theta_r)},$$

where

$$\theta_r = \beta_r d, \quad Z_{r,n} = \frac{\beta_r}{\beta'_r}$$

and

$$\gamma_{rn} = \frac{\beta_n I_{rn}^+}{\beta_n^2 - \beta_r^2}.$$

The perturbation theory just outlined assumed a guide closely fitted with a slug of ferrite of finite length, and slightly magnetized. The perturbation consisted in the small changes in permeability. Here we treat another kind of perturbation in which the ferrite does not fill the guide completely, and in which the magnetization can be arbitrarily large, but the sample is a thin lamina, mounted normally to the guide axis which only slightly perturbs the field pattern. Its shape can then be considered arbitrary; its thickness will be assumed uniform and very much smaller than a wavelength in the material.

Under these conditions, the equations for the perturbed amplitudes of the right-circularly polarized n^{th} radial mode are

$$\begin{aligned} \frac{\partial a_n}{\partial z} + j\beta_n b_n = j \frac{\omega}{\Delta_n} \left[\int (\mu - \mu_0) H_t \tilde{H}_{tn} dS - j \int \kappa H_t^* \tilde{H}_{tn} dS \right. \\ \left. + \int (\epsilon - \epsilon_0) E_z \cdot \tilde{E}_{zn} dS \right], \quad (30) \end{aligned}$$

$$\frac{\partial b_n}{\partial z} + j\beta_n a_n = j \frac{\omega}{\Delta_n} \left[\int (\epsilon - \epsilon_0) E_t \cdot \tilde{E}_{tn} dS \right],$$

in the usual notation. The terms on the right are assumed different from zero only in the range $z, z + \delta z$ occupied by the sample, and the integrals are extended over the cross section of the sample only. E_{tn}, H_{tn} are the mode functions for the empty guide. The fields H_t, E_t, E_z are the actual fields inside the sample. For the first-order calculation E_t and H_t are identical with the incident field and E_z (by continuity of the normal component of D_z) is $\frac{\epsilon_0}{\epsilon}$ times the incident E_z . Now the incident fields may

be taken simply as

$$a_{n0}E_{tn} ; \quad a_{n0}H_{tn} , \quad a_{n0}\bar{E}_{zn}$$

(since, *in vacuo*, $b_{n0} = a_{n0} \operatorname{sgn} \beta_n$ and $\operatorname{sgn} \beta_n = +1$ for the incident field), and the amplitude a_{n0} may be taken to be unity. Therefore equation (30) may be written

$$\frac{\partial a_n}{\partial z} = j \frac{\omega}{\Delta_n} \left[\int (\mu - \mu_0) H_{tn} \tilde{H}_{tn} dS - j \int \kappa H_{tn}^* \tilde{H}_{tn} dS \right. \\ \left. + \int \left(1 - \frac{\epsilon_0}{\epsilon} \right) \epsilon_0 E_{zn} \tilde{E}_{zn} dS \right] - j \beta_n b_n ,$$

$$\frac{\partial b_n}{\partial z} = j \frac{\omega}{\Delta_n} \int (\epsilon - \epsilon_0) E_{tn} \tilde{E}_{tn} dS - j \beta_n a_n .$$

We solve these equations by integrating through the small thickness of the sample. Since we are interested only in results of first order in δz , a_n and b_n on the right hand side of these equations may be replaced by a_{n0} , b_{n0} ; that is by 1, 1. Writing

$$I_A = \frac{\omega}{\Delta_n} \left[\int (\mu - \mu_0) H_{tn} \tilde{H}_{tn} dS - j \int \kappa H_{tn}^* \tilde{H}_{tn} dS \right. \\ \left. + \int \left(1 - \frac{\epsilon_0}{\epsilon} \right) \epsilon_0 E_{zn} \tilde{E}_{zn} dS \right] ,$$

$$I_B = \frac{\omega}{\Delta_n} \int (\epsilon - \epsilon_0) E_{tn} \tilde{E}_{tn} dS ,$$

we have, upon integration from the incidence plane (1) to the exit plane (2),

$$a_{n2} - a_{n1} = j(I_A - \beta_n) \delta z ,$$

$$b_{n2} - b_{n1} = j(I_B - \beta_n) \delta z .$$

a_{n1} , the amplitude at the entrance plane consists of $a_{n0} = 1$, the incident amplitude, plus a small reflected amplitude ρ . Thus $a_{n1} = 1 + \rho$. Similarly b_n consists of $a_{n0} \operatorname{sgn} \beta_n = 1$, and the small reflected amplitude $\rho \operatorname{sgn} \beta_n = -\rho$ (corresponding to a backward wave). b_{n2} is equal to a_{n2} (since the transmitted wave is a forward wave). Hence we have

$$a_{n2} - (1 + \rho) = j(I_A - \beta_n) \delta z ,$$

$$a_{n2} - (1 - \rho) = j(I_B - \beta_n) \delta z .$$

whose solutions for the transmitted and reflected amplitude are

$$a_{n2} = 1 - j \left(\beta_n - \frac{I_A + I_B}{2} \right) \delta z$$

and

$$\rho = j \frac{I_B - I_A}{2} \delta z.$$

Thus, to first order, the transmitted wave only undergoes a phase change, of amount

$$\delta\varphi_+ = \left(\beta_n - \frac{I_A + I_B}{2} \right) \delta z.$$

A similar result applies to a left circular wave; the only difference arises in the contribution

$$-\frac{\omega}{2\Delta_n} \int \kappa H_{tn}^* \cdot H_{tn} dS$$

to $\frac{I_A}{2}$ which will merely change sign. The remainder of I_A , and also I_B are unchanged. Thus the Faraday rotation of a plane polarized wave, which is one half the difference of the phase change for right and left circular waves, is equal to

$$\frac{\delta\varphi_+ - \delta\varphi_-}{2} = -\frac{\omega}{2\Delta_n} j \int_{\text{area of disc}} \kappa H_{tn}^* \tilde{H}_{tn} dS \delta z.$$

The integral on the right is real, and, when the disc is circular of radius r_1 , is of exactly the type that has already been encountered in the case of the ferrite embedded in a material with the same dielectric constant (Section 1.2). Here, however, κ need not be small, so long as δz is sufficiently small. In using the results of Section 1.2 it must be remembered that the mode functions there related to a dielectric, while here they relate to air-filled guide. When this is taken into account, one obtains for the specific rotation of a TE-mode:

$$\frac{\delta\varphi}{\delta z} = \frac{\beta_m}{(u_{nm}^2 - 1)J_n^2(u_{nm})} \left[\frac{n\kappa}{\mu_0} J_n^2 \left(u_{nm} \frac{r_1}{r_0} \right) \right], \quad (31a)$$

a TM-mode:

$$\frac{\delta\varphi}{\delta z} = \frac{\omega^2 \mu_0 \epsilon_0}{\beta_m j_{nm}^2 J_n'^2(j_{nm})} \left[\frac{n\kappa}{\mu_0} J_n^2 \left(j_{nm} \frac{r_1}{r_0} \right) \right], \quad (31b)$$

results which show that the rotation in a thin disc is independent of the dielectric constant of the disc.

2. THE PARALLEL PLANE CABLE

Measurements on ferrites are sometimes made by a coaxial cable technique. The ferrite fills a section of the cable, and is magnetized axially. In this section we consider a simplified problem of the same kind: that of wave motion between parallel conducting planes bounding a ferrite slab magnetized in the propagation direction. The characteristic equation for the propagation constant has been established by a number of authors, particularly Van Trier,¹ but as in the case of the cylindrical waveguide, no analysis of the equation in terms of the laboratory variables appears to have been made. We shall make such an analysis by means of the methods already used in Part I for the cylindrical waveguide. The discussion will be limited to the dominant (TEM-limit) mode; and to the principal features of the incipient modes (shape resonances).

The distance between the two conducting boundaries will be denoted by $2a$. The system of axes is as in Fig. 7; the y -axis is perpendicular to the walls (located at $y = \pm a$), the z -axis is the magnetization and propagation direction. The microwave fields are assumed not to vary along the x -axis.

The solution of Maxwell's equation proceeds just as in the case of the cylindrical waveguide (Part I, Section 3), with the simplifying feature that one of the coordinates (x) drops out of the problem. We shall use the reduced notation of Part I. β , the propagation constant will be measured in units of $\omega\sqrt{\mu_0\epsilon}$, the propagation constant of the unmagnetized, unbounded medium. $\bar{a}$ will denote $a\omega\sqrt{\mu_0\epsilon}$, the half-spacing measured in terms of the reduced wavelength $\frac{\lambda}{2\pi}$ of the unmagnetized, extended medium.

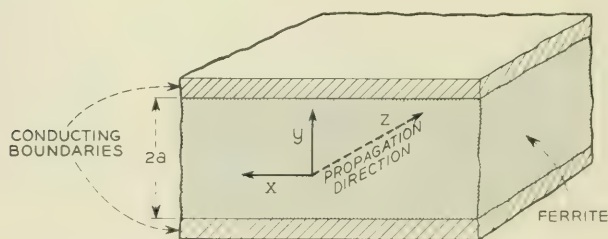


Fig. 7 — Parallel plane cable filled with ferrite magnetized along propagation direction.

It is found that (see Part I, Section 3) the longitudinal electric and magnetic fields are then given by

$$E_z = \frac{\Lambda_2 \psi_1 - \Lambda_1 \psi_2}{\Lambda_2 - \Lambda_1}; \quad H_z = j \frac{\psi_1 - \psi_2}{\Lambda_2 - \Lambda_1},$$

where $\psi_{1,2}(y)$ are solutions of

$$\frac{\partial \psi_{1,2}}{\partial y^2} + \chi_{1,2}^2 \psi_{1,2} = 0, \quad (32)$$

where $\Lambda_{1,2}$ are the roots of the quadratic (16), Part I, and where, finally, the χ 's are defined in terms of the Λ by equations (17a) and (17b), Part I. The x -component of E is given by [see equation (22a), Part I]

$$(\Lambda_1 - \Lambda_2) \Omega E_x = \left[\beta \Lambda_1 \rho_H - \rho_H^2 \nu_H + \nu_H \left(1 - \frac{\beta^2}{\nu_H} \right) \right] \frac{\partial \psi_2}{\partial y},$$

or the same expression with suffixes 1, 2 interchanged. (Note that the x component of $\nabla \psi$ is now zero, that of $\nabla^* \psi$ is $\frac{\partial \psi}{\partial y}$). If the boundary conditions are to be satisfied at both $y = +a$ and $y = -a$, the solutions of equation (32) must be either even or odd functions of y . Since E_z must vanish at $y = \pm a$, we have for the symmetric (or even) case

$$\psi_1^s = \Lambda_1 \frac{\cos \chi_1 y}{\cos \chi_1 a}; \quad \psi_2^s = \Lambda_2 \frac{\cos \chi_2 y}{\cos \chi_2 a},$$

and for the antisymmetric (odd) case

$$\psi_1^a = \Lambda_1 \frac{\sin \chi_1 y}{\sin \chi_1 a}; \quad \psi_2^a = \Lambda_2 \frac{\sin \chi_2 y}{\sin \chi_2 a}.$$

The characteristic equation for these two sets now follows from the fact that $E_x = 0$ at $y = \pm a$. Expressing the Λ 's in terms of the χ 's by means of equation (17b), Part I, and writing $\beta \Lambda_{2,1} = \lambda_{1,2}$, we obtain

$$\frac{1}{\lambda_1 \chi_1^2} \chi_1 \bar{a} \tan \chi_1 \bar{a} = \frac{1}{\lambda_2 \chi_2^2} \chi_2 \bar{a} \tan \chi_2 \bar{a}$$

for the symmetric case, and

$$\lambda_1 \chi_1^2 \frac{\tan \chi_1 \bar{a}}{\chi_1 \bar{a}} = \lambda_2 \chi_2^2 \frac{\tan \chi_2 \bar{a}}{\chi_2 \bar{a}}$$

for the antisymmetric case. To bring these equations into a form suitable for graphical discussion, we express the χ in terms of λ , σ by means

of the quadratic for λ , and by means of the relations between the tensor components of the permeability (see beginning of Section 4.11, Part I). We then obtain

$$\begin{aligned} \frac{1}{\lambda_1} \sqrt{\frac{1 - \sigma\lambda_1}{1 - \lambda_1^2}} \tan \bar{a} \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} \\ = \frac{1}{\lambda_2} \sqrt{\frac{1 - \sigma\lambda_2}{1 - \lambda_2^2}} \tan \bar{a} \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \end{aligned} \quad (33)$$

for the symmetric modes, and

$$\lambda_1 \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} \tan \bar{a} \sqrt{\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}} = \lambda_2 \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \tan \bar{a} \sqrt{\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}} \quad (34)$$

for the antisymmetric modes. When λ_1, λ_2 pairs satisfying one of these equations, and the Polder relation

$$\lambda_1 + \lambda_2 - \sigma\lambda_1\lambda_2 = \sigma + p \quad (35)$$

have been found, β^2 is given by

$$\beta^2 = -\lambda_1\lambda_2.$$

Appearances to the contrary, equations (33) and (34) describe the ordinary TE and TM modes when $p = 0$, regardless of σ (see the discussion opposite Fig. 3. of Part I). When $\sigma = 0$, the Polder relation transforms $\frac{1 - \lambda_2^2}{1 - \sigma\lambda_2}$ into $\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1}$, so that either of equations (33) and (34) can be satisfied only by

$$\frac{1 - \lambda_1^2}{1 - \sigma\lambda_1} = \frac{1 - \lambda_2^2}{1 - \sigma\lambda_2} = n^2 \frac{\pi^2}{\bar{a}^2} \quad \text{or} \quad \left(n + \frac{1}{2}\right)^2 \frac{\pi^2}{\bar{a}^2}$$

and the two alternatives respectively give

$$-\lambda_1\lambda_2 = \beta^2 = 1 - n^2 \frac{\pi^2}{\bar{a}^2}$$

or

$$1 - \left(n + \frac{1}{2}\right)^2 \frac{\pi^2}{\bar{a}^2}.$$

Since both equations (33) and (34) are satisfied by either the n -, or the $(n + \frac{1}{2})$ -dependence, a problem of classification arises: "Which mode, TE or TM, is described, as $p \sim 0$, by the solutions of (33) and (34)?" Leaving aside the question of the TEM mode we note that the following degeneracy exists in the isotropic case:

A TE mode with antisymmetric H_z and a TM mode with symmetric E_z have the same propagation constant

$$\beta^2 = 1 - \frac{\pi^2}{\bar{a}^2} \left(n + \frac{1}{2} \right)^2 \quad (n = 0, 1, 2, \dots).$$

A TE mode with symmetric H_z and a TM mode with antisymmetric E_z have the same propagation constant

$$\beta^2 = 1 - \frac{n^2 \pi^2}{\bar{a}^2} \quad (n = 1, 2, \dots).$$

Since application of a very small magnetization will not destroy the symmetry properties, it follows that the solution of equation (34), and the solution of equation (33), which in the limit $p = 0$ reduce to

$$\beta^2 = 1 - \frac{\pi^2}{\bar{a}^2} \left(n + \frac{1}{2} \right)^2,$$

correspond to a TE-limit and a TM-limit mode respectively. Likewise, the solutions of equation (33) and the solution of equation (34) which in the limit $p = 0$ reduce to

$$\beta^2 = 1 - n^2 \frac{\pi^2}{\bar{a}^2}$$

correspond to a TE-limit and a TM-limit mode respectively. When $p \neq 0$, the new β^2 are thus given by different equations, so that the magnetization is seen to have removed the degeneracy of the isotropic case.

In the special case of the TEM limit mode ($\beta^2 = 1$ when $p = 0$) only one of equations (33) and (34) has real roots. For, when $p = 0$, the Polder relation gives

$$\begin{aligned} \frac{1 - \lambda_1^2}{1 - \sigma \lambda_1} &= \frac{1 - \lambda_2^2}{1 - \sigma \lambda_2}, \\ &= 0, \end{aligned} \quad (36)$$

if β^2 is to equal unity. But equation (36) satisfies equation (34) only [equation (33) would demand $\lambda_1 = \lambda_2$ which is impossible for real β]. Thus the TEM mode exists as the limit of an antisymmetric mode only.

In fact it is easily shown that for a value of $\bar{a}$ too small (less than $\frac{\pi}{2}$) to admit any except the TEM mode in the isotropic medium, the only solutions of equation (33) for general σ , p describe incipient modes.

Thus for $\bar{a} < \frac{\pi}{2}$ the analysis may be confined to equation (34) which will give the course of the TEM limit mode. The graphical analysis follows closely that of the cylindrical waveguide (Part I, Section 4.1); the part played by the G -function there is now played by the function

$$L = \lambda \sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda_1}} \tan \bar{a} \sqrt{\frac{1 - \lambda^2}{1 - \sigma\lambda}}.$$

A contour map is drawn of the function L in the upper half of the λ - σ plane. Only the upper half plane need be considered since the L -equation (34) is unchanged by reflection in the origin. This, incidentally, means that the present situation is reciprocal, so that σ , p hereafter can be considered positive. Pairs of values $\lambda_1(>0)$ and $\lambda_2(<0)$ satisfying the L -equation can then immediately be read off, but while they will give $\beta^2 = \lambda_1\lambda_2$, they do not necessarily, for given σ , p , satisfy the Polder relation, equation (35). To ensure this, the quadrant, $\lambda < 0$, $\sigma > 0$, and the L -contours therein are transformed on to $\lambda > 0$, $\sigma > 0$ by the Polder relation for fixed p :

$$\lambda_2 = \frac{\sigma + p - \lambda_1}{1 - \sigma\lambda_1} = T(\lambda_1).$$

The surfaces $L_1 = L(\lambda_1, \sigma)$ and $L_2 = L(T(\lambda_1), \sigma)$ will intersect in a number of curves, along whose base curves in the λ - σ plane both the Polder equation and the L equation are satisfied, and along which $\beta^2 = -\lambda_1\lambda_2$ is known as a function of σ , p . The zero and infinity curves of $L(\lambda, \sigma)$ are denoted by 0, I when $\lambda > 0$, and by $0'$, I' when $\lambda < 0$. The transforms of $0'$, I' onto $\lambda > 0$ are denoted by $(0')_T$, $(I')_T$. The suffix n denotes the infinity (zero) curves corresponding to

$$\frac{1 - \lambda^2}{1 - \sigma\lambda} = \left(n + \frac{1}{2}\right)^2 \frac{\pi^2}{\bar{a}^2}; \quad \left(\frac{1 - \lambda^2}{1 - \sigma\lambda} = n^2 \frac{\pi^2}{\bar{a}^2}\right) \quad n \text{ integral.}$$

The lines $\lambda = 0, +1, -1$, are all zero curves denoted by $0_A, 0_B, 0_{B'}$, respectively. The line $\lambda\sigma = 1$ is a conditional infinity curve, called I_c . It is an I curve when viewed from $\sigma\lambda > 1$ for $\lambda < 1$, and when viewed from $\sigma\lambda < 1$ for $\lambda > 1$. Otherwise (for $\sigma\lambda < 1, \lambda < 1$ and $\sigma\lambda > 1, \lambda > 1$) it is a limit curve of all possible curves $L = \text{const.}$, where the constant takes on any value indefinitely many times. (See Part I, Section 4, where the curve $\sigma\lambda = 1$ is a conditional zero curve, 0_c .) Fig. 8, drawn for $\bar{a} \sim 1$, $p \sim 0.5$, shows the part of the first quadrant allowed by the Polder relation divided into regions of like and unlike sign of $L(\lambda, \sigma)$ and $L(T(\lambda_1, \sigma))$ by the various 0, I and $(0')_T$, $(I')_T$ curves. The un-

shaded regions are regions of like sign, and all these carry solution curves (dotted lines), by the same reasoning as was employed in Part I, Section 4.11. Two branches of the TEM limit mode exist; in the area bounded by $(0_A)_T$, 0_B , $(0_B')_T$ and in the area 0_B , I_1 , $(0_B')_T$. The branch in the first region begins at $\sigma = 0$ with $\lambda_1\lambda_2$ given by

$$\lambda_1\sqrt{1-\lambda_1^2}\tan\bar{a}\sqrt{1-\lambda_1^2}=\lambda_2\sqrt{1-\lambda_2^2}\tan\bar{a}\sqrt{1-\lambda_2^2},$$
$$\lambda_1+\lambda_2=p$$

and ends at the intersection of $(0_A)_T$, 0_B with $\sigma = 1 - p$ and $\beta^2 = 0$ (since $\lambda_2 = 0$ on $(0_A)_T$). The branch in the region 0_B , I_1 , $(0_B')_T$ begins with $\beta_2 = \infty$, $\sigma = 1$ and proceeds towards $\sigma = \infty$, $\beta^2 = 1$.

The region bounded by I_1 , $(0_A)_T$, I_c contains an infinity of solution curves, the incipient modes. The n^{th} of these begins at $\sigma = 1$, $\lambda_1 = 1$, the intersection of I_n and I_c (the line I_c is also the transform of $\lambda_2 = -\infty$,

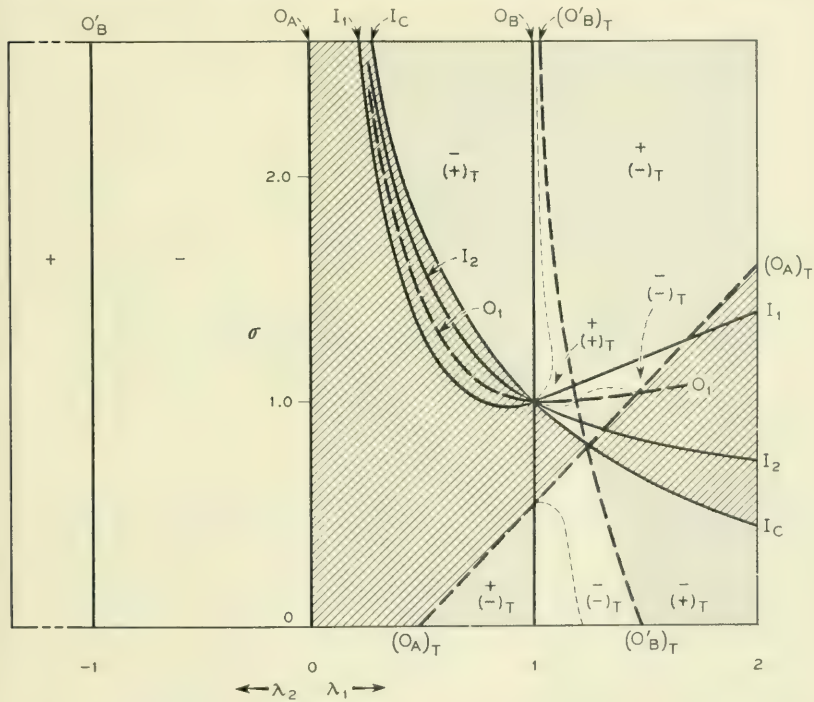


Fig. 8 — Zero and infinity contours of $L(\lambda_1, \sigma)$ and $L(T(\lambda_1), \sigma)$ in the first quadrant of the $\lambda - \sigma$ plane. Cross hatched regions are excluded by the Polder relation; shaded areas are regions of unlike sign of the two functions. Dotted curves are solution curves.

along which $L_2 = \infty$); crosses from one allowed region to the next through the intersection of 0_n with $(0_B')_T$ and proceeds to its cut-off point $(0_n - (0_A)_T)$.

Formulae giving β^2 in the neighborhood of special points on the solution curves, or for special values of the parameters, are obtained exactly as in Part I, by expansion of the L -equation. The results are tabulated below. They relate to antisymmetric modes only.

Cut-off of the TEM-limit mode ($\sigma = 1 - p, \beta^2 = 0$)

Near cut-off

$$\frac{d\beta^2}{d\sigma} = -\frac{1}{p \left(1 - \frac{\tan \bar{a}}{2\bar{a}}\right)} \quad (37)$$

For $p > 1$, the TEM mode has no lower branch.

Behavior near resonance ($\sigma = 1$):

$$\begin{aligned} \text{TEM mode: } \beta^2 &= \frac{\frac{\pi^2}{8\bar{a}^2}}{1 - \sigma} p, \\ n^{\text{th}} \text{ incipient mode: } \beta^2 &= \left(\frac{\left(n + \frac{1}{2}\right)^2 \pi^2}{2\bar{a}^2} - 1 \right) \frac{p}{1 - \sigma}. \end{aligned} \quad (38)$$

($n = 1, 2, \dots$)

Cut-off of the n^{th} incipient mode (parametric representation):

$$\begin{aligned} \lambda_1 &= e^\theta; \quad \sigma = e^{-\theta} \left(1 - \frac{\bar{a}^2}{n^2 \pi^2} \right) + \frac{\bar{a}^2}{n^2 \pi^2} e^\theta \\ p &= \sigma - \lambda_1; \quad \beta^2 = 0 \\ \frac{d\beta^2}{d\sigma} &= \frac{m^2 \pi^2}{\bar{a}^2} \left(\frac{m^2 \pi^2}{\bar{a}^2} - 1 \right) \coth \theta \bigg/ \left[\left(\frac{m^2 \pi^2}{\bar{a}^2} - 1 \right) e^{-\theta} + \frac{2 \tan \bar{a}}{\bar{a}} e^{-\theta} - e^\theta \right]. \end{aligned} \quad (39)$$

(The reader may note the similarities of these formulas to those for the cylindrical waveguide).

Spot-point for the n^{th} incipient mode:

$$\begin{aligned} \sigma &= \left[\left(1 - \frac{\bar{a}^2}{n^2 \pi^2} \right) \bigg/ \beta^2 \right] + \frac{\bar{a}^2}{n^2 \pi^2} \beta^2, \\ p &= (1 - \sigma) (\beta^2 - 1). \end{aligned} \quad (40)$$

Approximate formulae for all antisymmetric modes at small p , $\sigma \neq 1$:

$$\text{TM modes: } \beta^2 = \beta_{n,1}^2 \frac{\sigma p}{1 - \sigma^2}, \quad (41)$$

$$\text{TE modes: } \beta^2 = \beta_{n,2}^2 \left(1 - \frac{\sigma p}{1 - \sigma^2} \right), \quad (42)$$

where

$$\beta_{n,1}^2 = 1 - \frac{n^2 \pi^2}{\bar{a}^2} \quad n = 1, 2, \dots$$

and

$$\beta_{n,2}^2 = 1 - \frac{\left(n + \frac{1}{2}\right)^2 \pi^2}{\bar{a}^2} \quad n = 0, 1, 2, \dots$$

TEM mode:

$$\beta^2 = 1 - \frac{\sigma p}{1 - \sigma^2}. \quad (43)$$

TEM mode for large σ :

$$\beta^2 = 1 + \frac{p}{\sigma} + O\left(\frac{1}{\sigma^3}\right). \quad (44)$$

TEM mode for very small $\bar{a}$:

for sufficiently small $\bar{a}$, the β^2 for the TEM mode is independent of $\bar{a}$. For general σ , p , except $\sigma = 1$,

$$\beta^2 = \frac{1 - (\sigma + p)^2}{1 - \sigma p - \sigma^2}. \quad (45)$$

Reference to the expression for μ and κ in terms of σ , p will show that the wave progresses as though the medium had a scalar permeability μ_{eff} given by

$$\frac{1}{\mu_{\text{eff}}} = \frac{1}{2} \left(\frac{1}{\mu + \kappa} + \frac{1}{\mu - \kappa} \right).$$

3. FIELD PATTERNS IN THE FILLED CIRCULAR GUIDE

In Part I we discussed propagation in a circular guide filled with ferrite and longitudinally magnetized. Formulae for the field com-

ponents were given there, and it is of some interest to examine numerically or graphically the distribution of E and H . Since all field components vary (in complex notation) as $e^{j(n\theta + \omega t)}$, with $n = 1$ in the present case, it is clear that the field patterns are stationary in a system rotating with angular velocity, $-\omega$. Writing Φ for the angular variable, $\theta + \omega t$, in this system, one finds the following formulas:

$$\begin{aligned} E_x &= -E_- \sin 2\Phi, & E_r &= (E_+ - E_-) \sin \Phi, \\ E_y &= E_- \cos 2\Phi + E_+, & E_\theta &= (E_+ + E_-) \cos \Phi, \\ E_z &= E_0 \cos \Phi, \\ H_x &= H_- \cos 2\Phi + H_+, & H_r &= (H_+ + H_-) \cos \Phi, \\ H_y &= H_- \sin 2\Phi, & H_\theta &= (H_- - H_+) \sin \Phi, \\ H_z &= H_0 \sin \Phi, \end{aligned} \tag{46}$$

where

$$\begin{aligned} E_-(r) &= \frac{\beta \left(\frac{1}{\lambda_1} + 1 \right)}{2\chi_1 J_1(\chi_1 r_0)} J_2(\chi_1 r) - \left[\quad \right]_2, \\ E_+(r) &= \frac{\beta \left(1 - \frac{1}{\lambda_1} \right)}{2\chi_1 J_1(\chi_1 r_0)} J_0(\chi_1 r) - \left[\quad \right]_2, \\ H_-(r) &= -\frac{1 + \frac{1 - \chi_1^2}{\lambda_1}}{2\chi_1 J_1(\chi_1 r_0)} J_2(\chi_1 r) - \left[\quad \right]_2, \\ H_+(r) &= \frac{-1 + \frac{1 - \chi_1^2}{\lambda_1}}{2\chi_1 J_1(\chi_1 r_0)} J_0(\chi_1 r) - \left[\quad \right]_2, \\ E_0(r) &= \frac{J_1(\chi_1 r)}{J_1(\chi_1 r_0)} - \left[\quad \right]_2, \\ H_0(r) &= -\frac{1}{\lambda_1} \frac{J_1(\chi_1 r)}{J_1(\chi_1 r_0)} - \left[\quad \right]_2. \end{aligned} \tag{47}$$

The terms in square brackets are in each case the same as the corresponding unbracketed terms, λ_2 and χ_2 replacing λ_1 and χ_1 . The quantities E_+ and E_- are the amplitudes of the left-handed and right-handed components of circular polarization into which the transverse E -field may be resolved at each point. H_+ and H_- have a similar significance

for the transverse H -field. This may be readily verified by examining the vectors $(E_x + jE_y)e^{-j\omega t}$ and $(H_x + jH_y)e^{j\omega t}$, which represent the transverse field vectors in the laboratory system. The transverse fields are elliptically polarized at any point and the ratio of minor to major axes of the ellipses are

$$\frac{||E_+|| - ||E_-||}{||E_+|| + ||E_-||} \quad \text{and} \quad \frac{||H_+|| - ||H_-||}{||H_+|| + ||H_-||}.$$

The fields so far are normalized only by the choice of a simple form for the function, $E_0(r)$; all components may, of course, be multiplied by the same arbitrary constant. There is some virtue to a normalization based upon power flow. The power flow is given in unreduced units by

$$\int_{\text{guide}} (E \times H)_z dS.$$

Using the scaled units of part I with

$$r_{\text{actual}} = \frac{r}{\omega \sqrt{\mu_0 \epsilon}}$$

and H replaced by $\sqrt{\frac{\mu_0}{\epsilon}} H$ to give it the same dimensions as E (ϵ is the dielectric constant of the ferrite), the power flow becomes in the present variables

$$\begin{aligned} P &= -\sqrt{\frac{\epsilon}{\mu_0}} \frac{\pi}{\omega^2 \mu_0 \epsilon} \int (E_+ H_+ + E_- H_-) r dr, \\ &= -\sqrt{\frac{\epsilon}{\mu_0}} \frac{\pi}{\omega^2 \mu_0 \epsilon} I. \end{aligned}$$

We shall normalize the E and H fields by dividing the values given by equations (47) by $I^{1/2}$. This makes the power per (isotropic wavelength in the ferrite)² a constant. In Fig. 9 we show the normalized fields for the TE_{11} -limit and TM_{11} -limit modes as a function of r for the case $r_0 = 5.75$, $|p| = 0.6$ and several values of σ . The behavior of the modes as a function of σ and p for this radius is shown in Fig. 14 of Part I. We also show the amplitudes for the isotropic cases, $\sigma = p = 0$. It may be recalled from the discussion of cut-off points in Part I that, for the TM mode at this radius, when cut-off is reached at $\sigma = -0.4$, the amplitude of the H^+ field is overwhelmingly greater than that of the others. Even when normalized to the same power flow, the field amplitudes for a given σ are undetermined to a factor of ± 1 . This factor has been

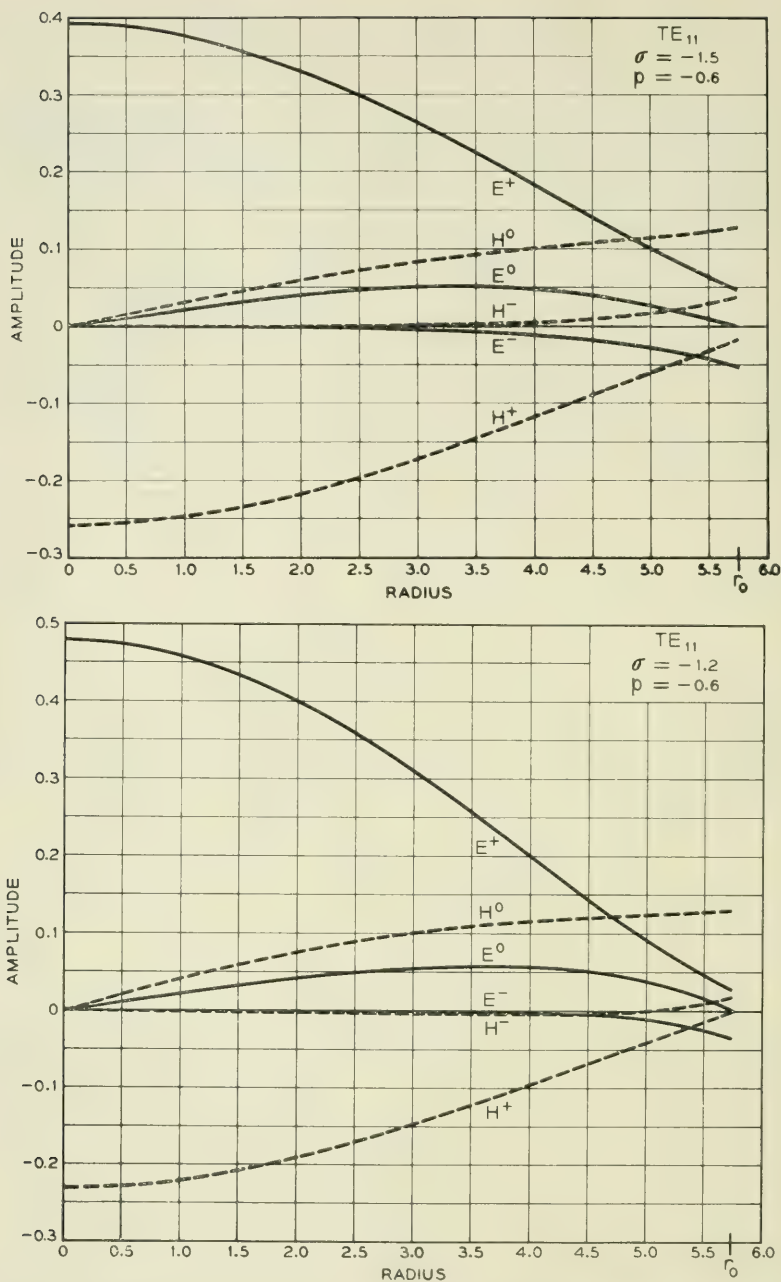


Fig. 9 — Amplitude of the field components in filled cylindrical guide, longitudinally magnetized, as a function of radius, for the TE_{11} -limit modes at several values of σ . $r_0 = 5.75$. $|p| = 0.6$, except for Figs. 9(f) and (p) which are the isotropic cases, $\sigma = p = 0$. The fields are normalized to a constant power flow per (free space wavelength)². (a), top, (b), bottom.

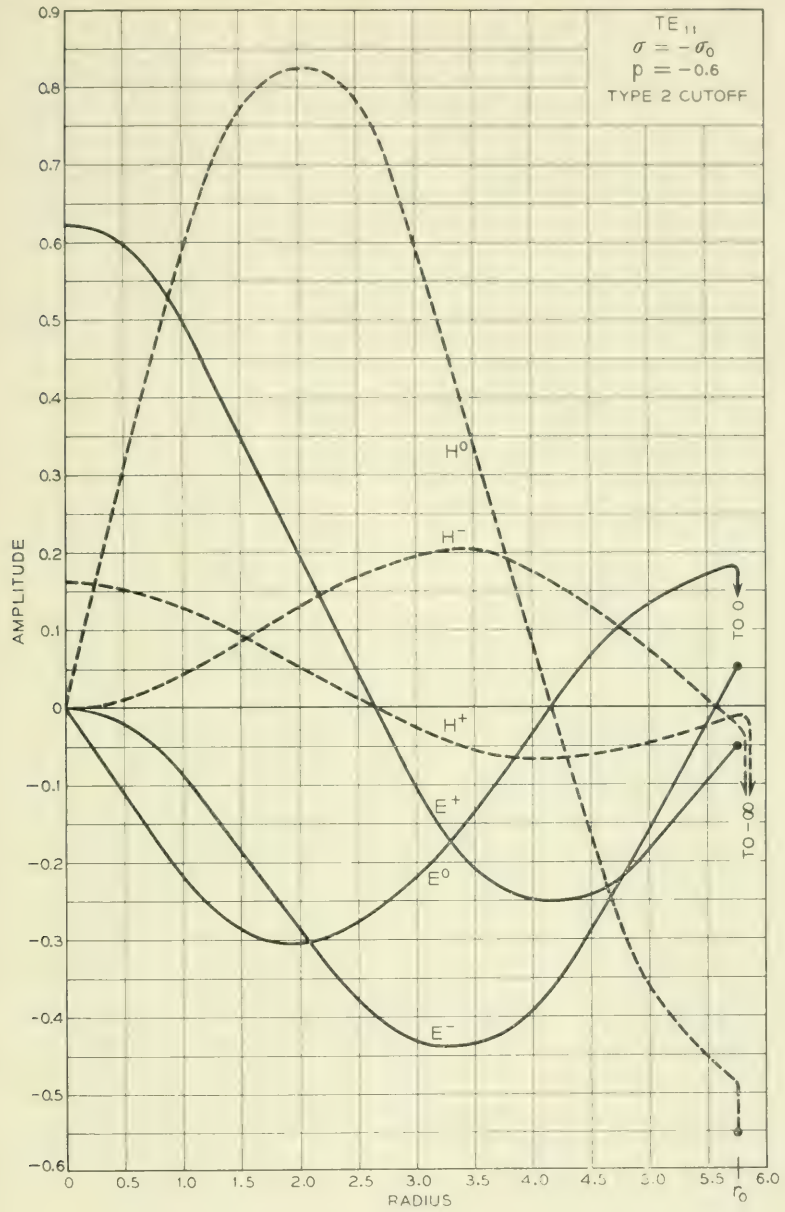


Fig. 9(c) — See Fig. 9.

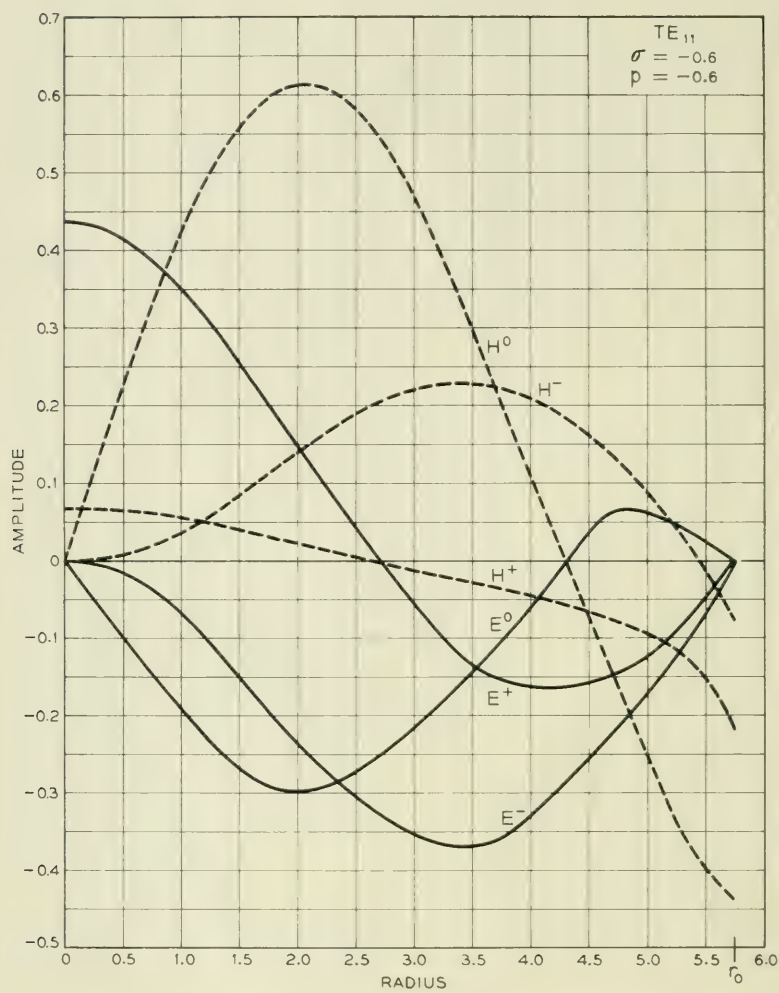


Fig. 9(d) — See Fig. 9.

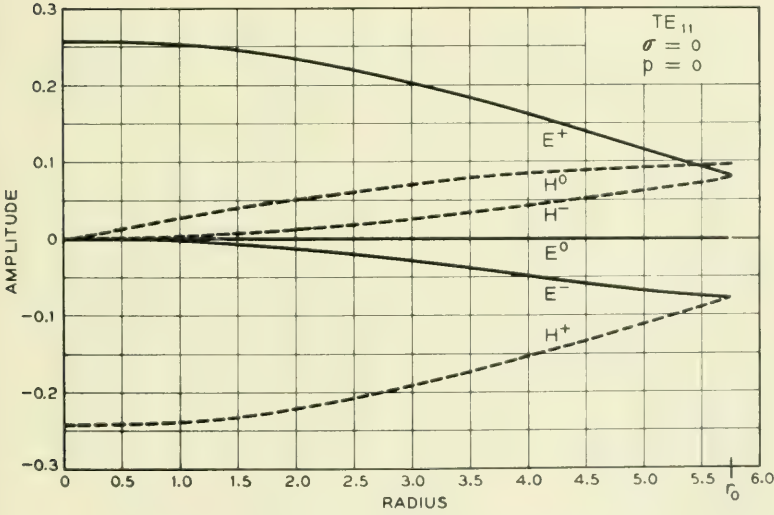
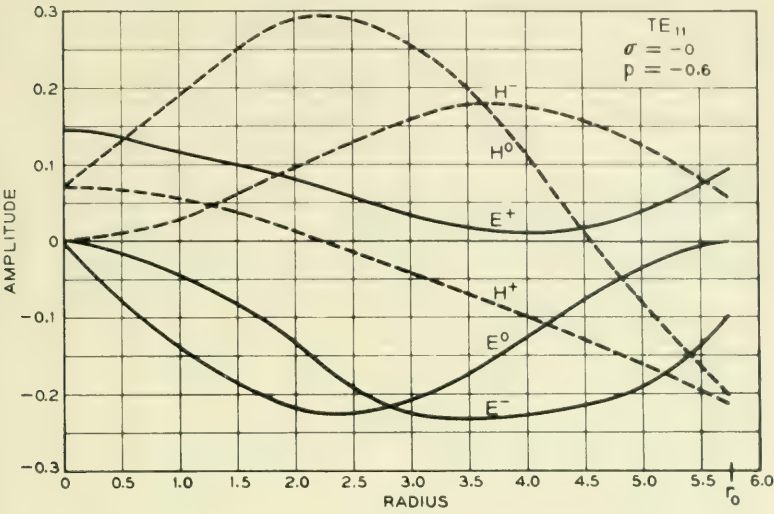


Fig. 9(e), top, and (f), bottom — See Fig. 9.

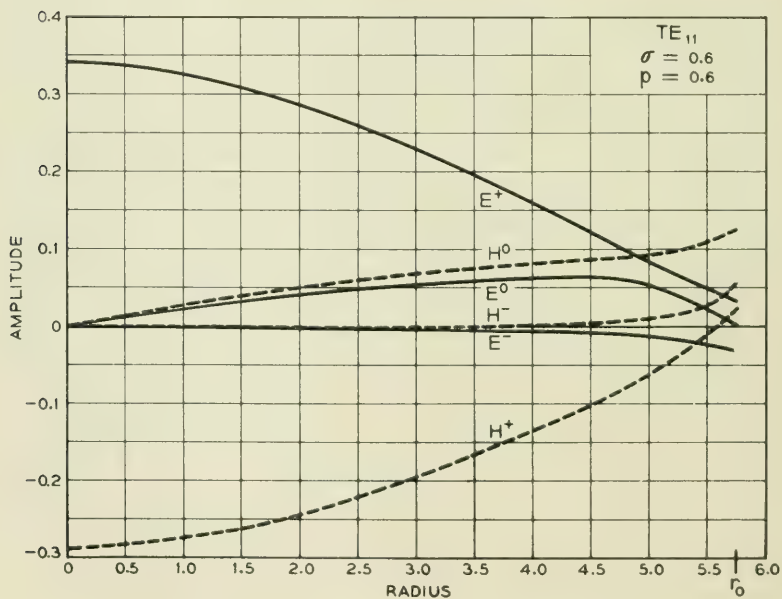
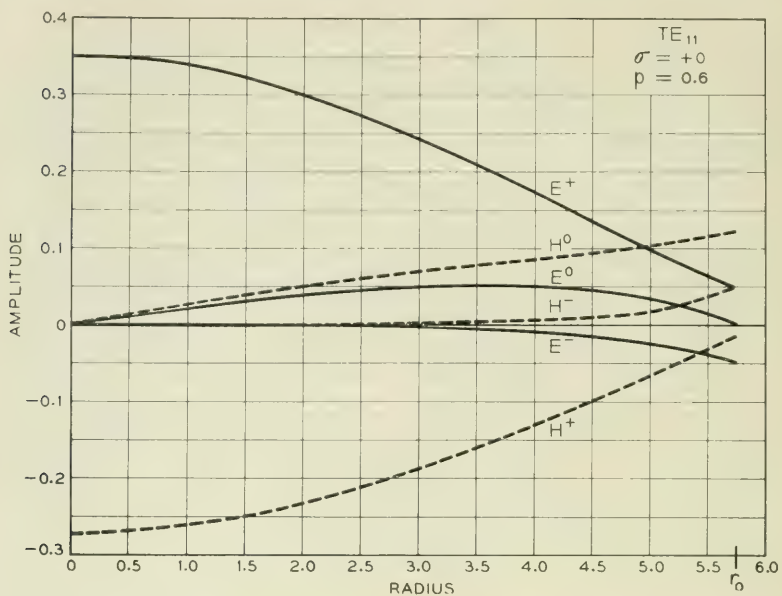


Fig. 9(g), top, and (h), bottom — See Fig. 9.

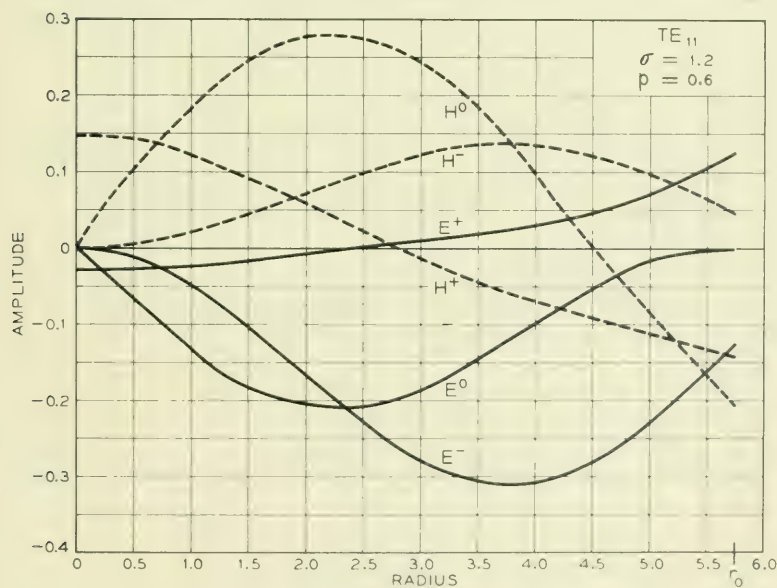
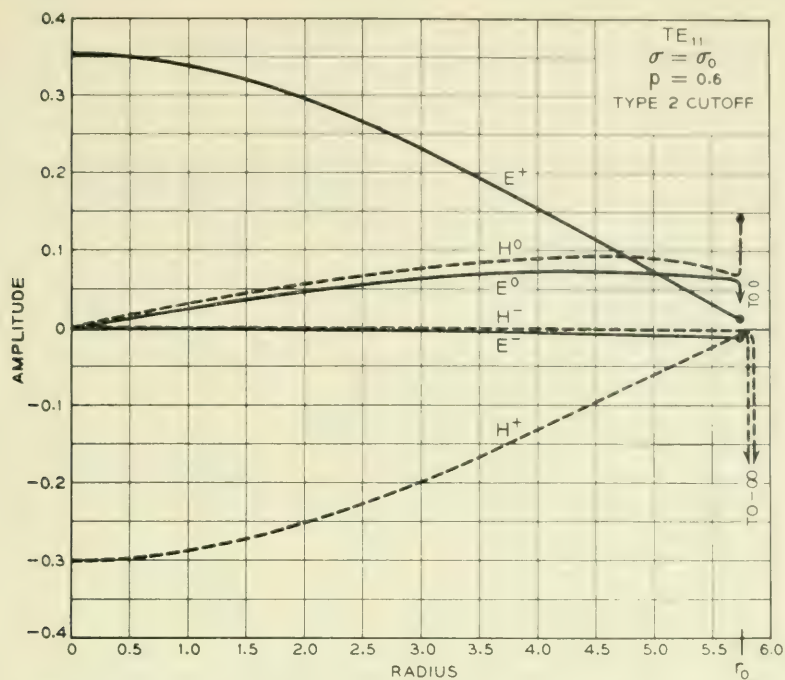


Fig. 9(i), top, and (j), bottom — See Fig. 9.

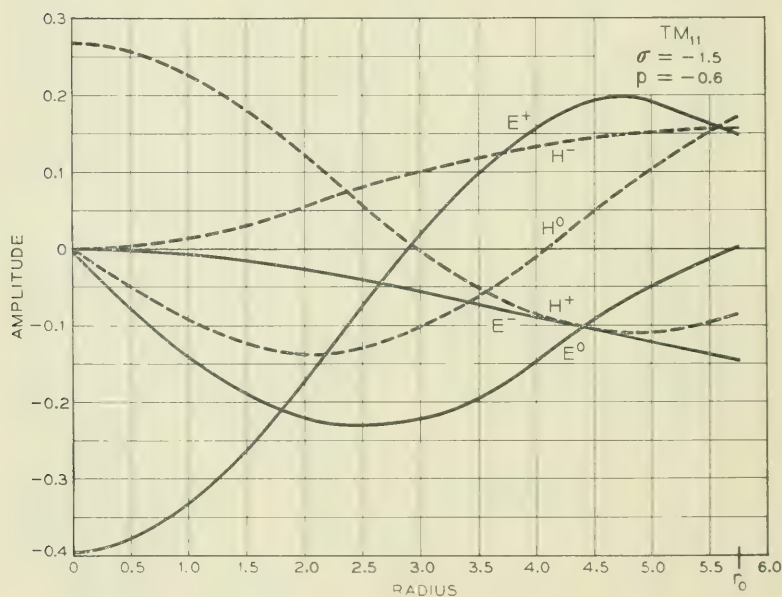
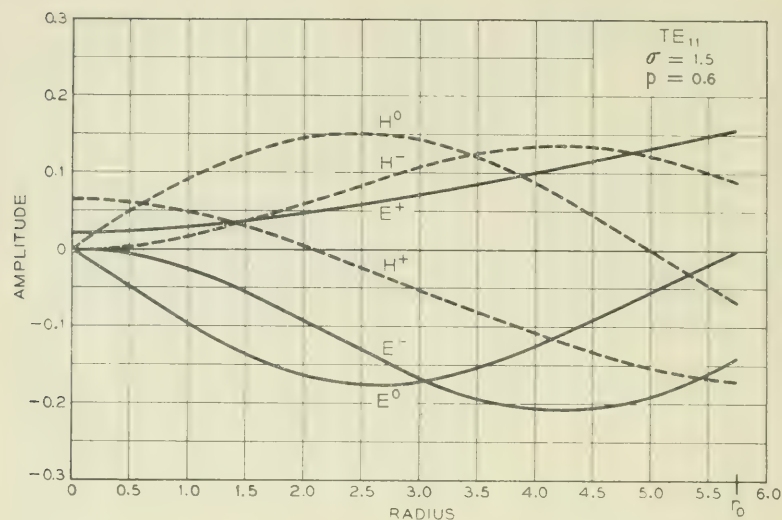


Fig. 9(k), top, and (l), bottom — See Fig. 9.

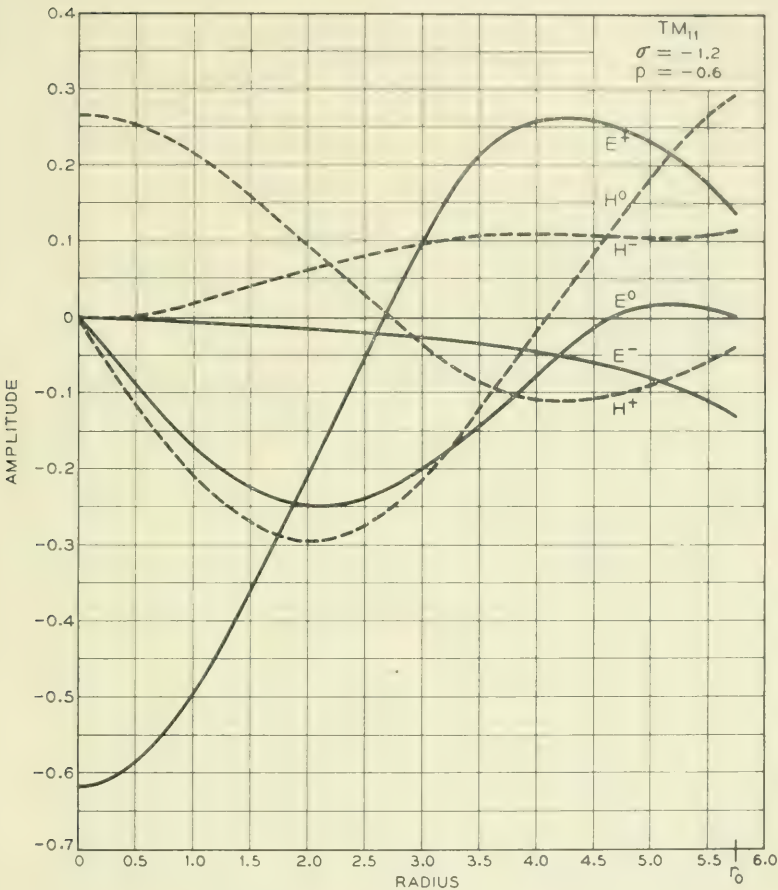


Fig. 9(m) — See Fig. 9.

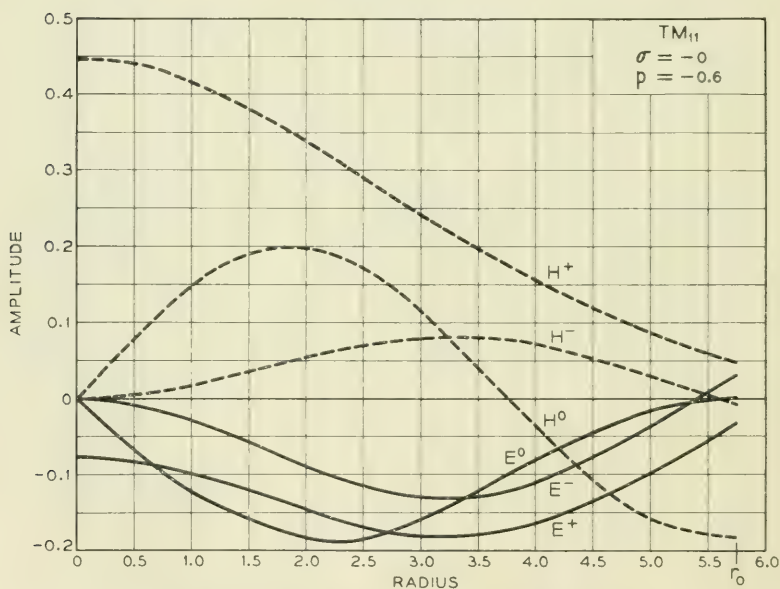
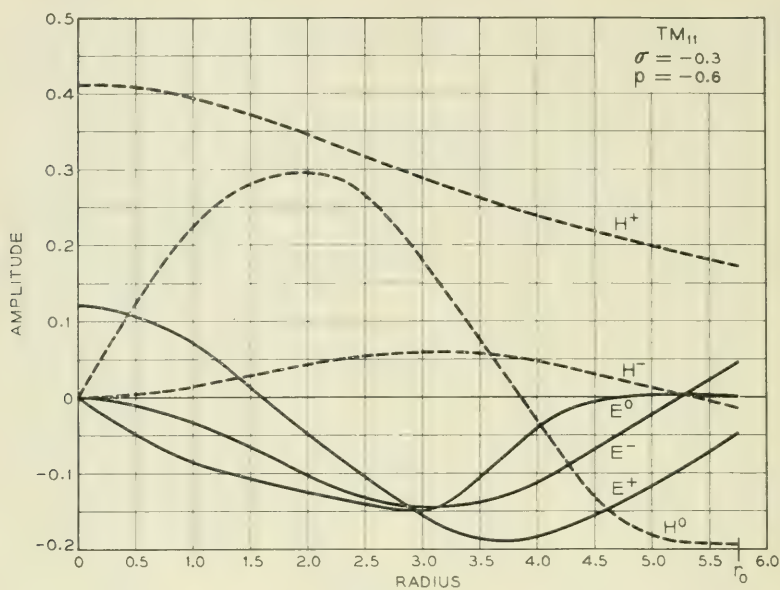


Fig. 9(n), top, and (o), bottom See Fig. 9.

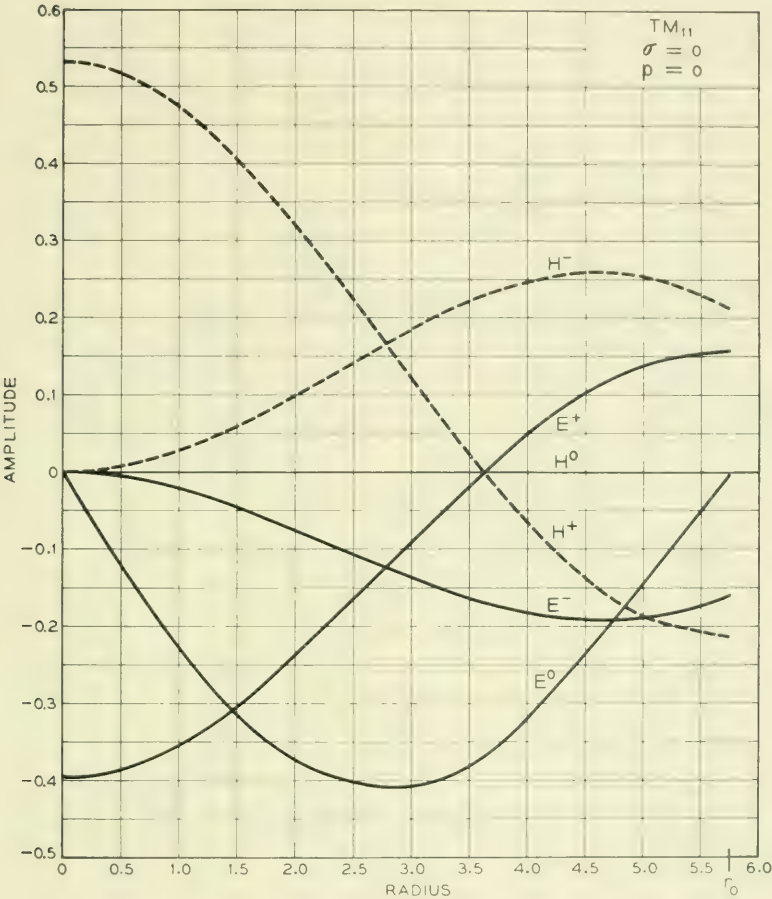


Fig. 9(p) — See Fig. 9.

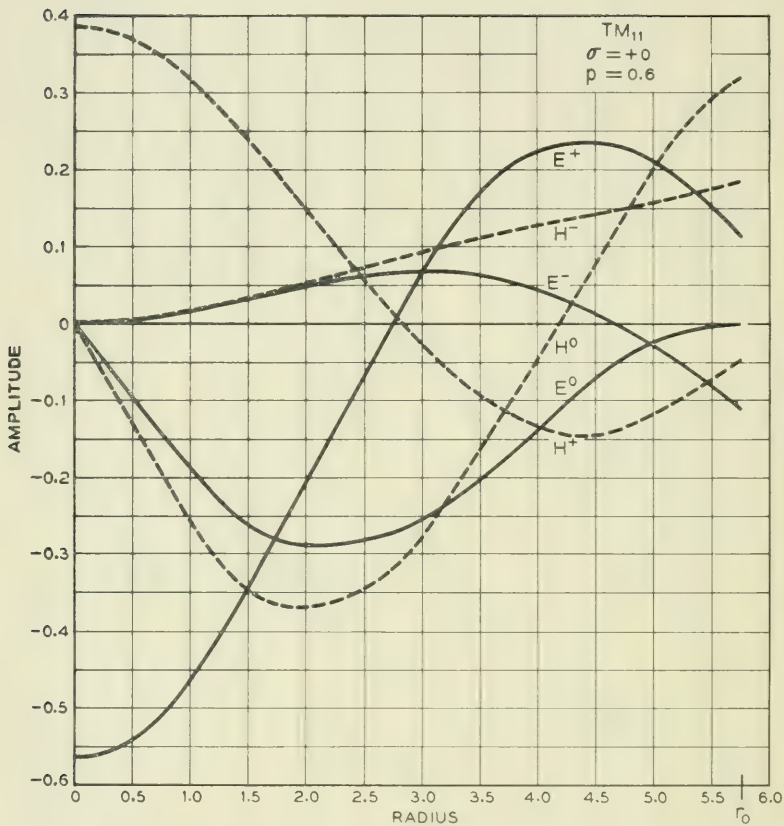


Fig. 9(q) — See Fig. 9.

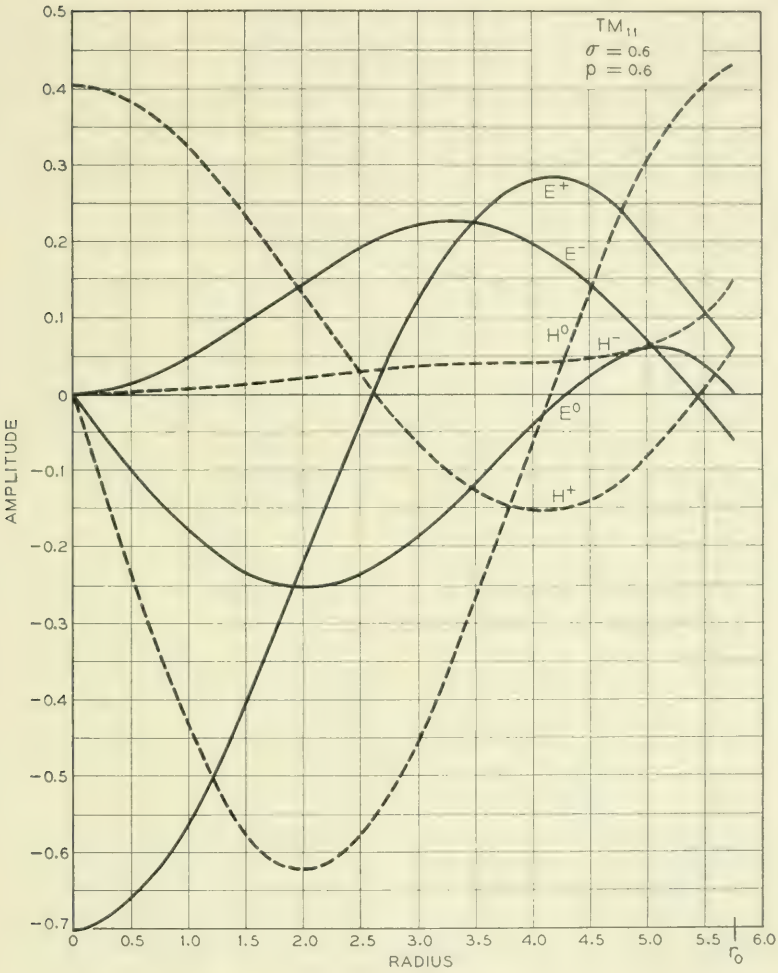


Fig. 9(r) — See Fig. 9.

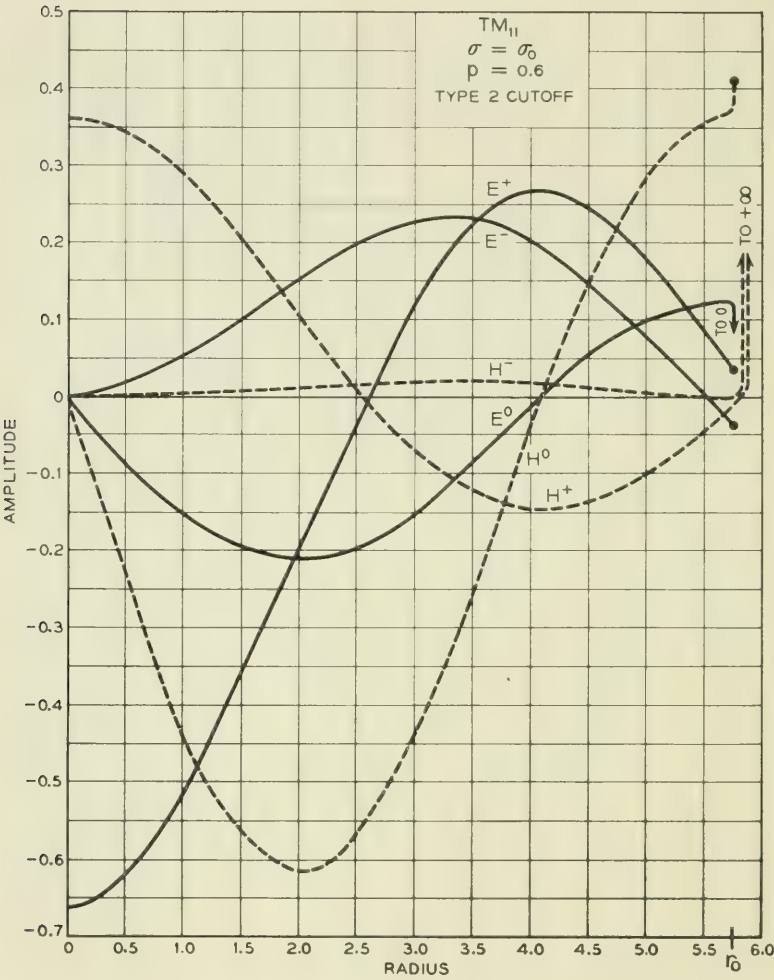


Fig. 9(s) — See Fig. 9.

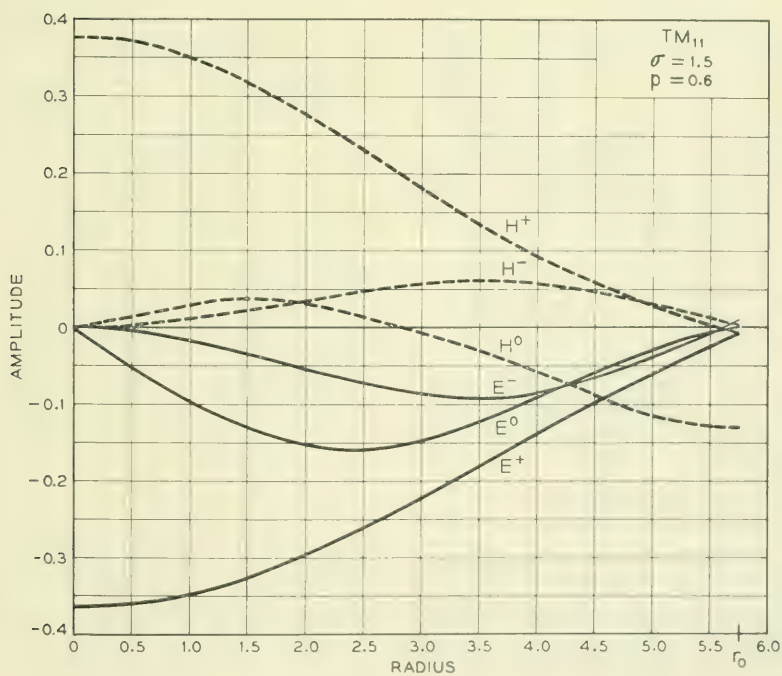
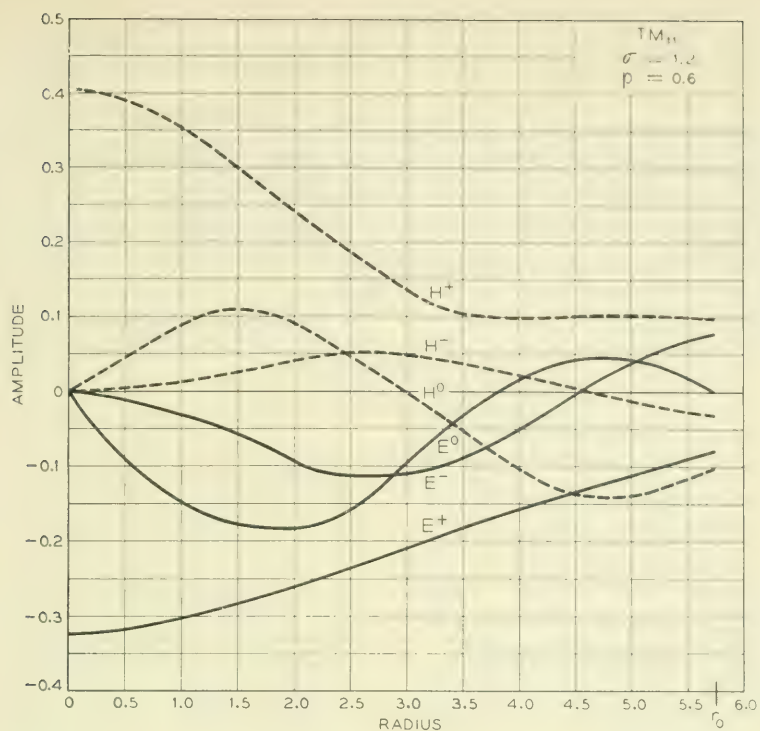


Fig. 9(t), top, and (u), bottom — See Fig. 9.

chosen for each pattern according to the following considerations. For large $|\sigma|$ the amplitudes should tend to those of the isotropic case; for $|\sigma| < \sigma_0$ the amplitudes should develop with increasing $|\sigma|$ away from the isotropic amplitudes. Thus, if the phase of the isotropic amplitudes is first fixed, that of all the others may be fixed by the conditions that

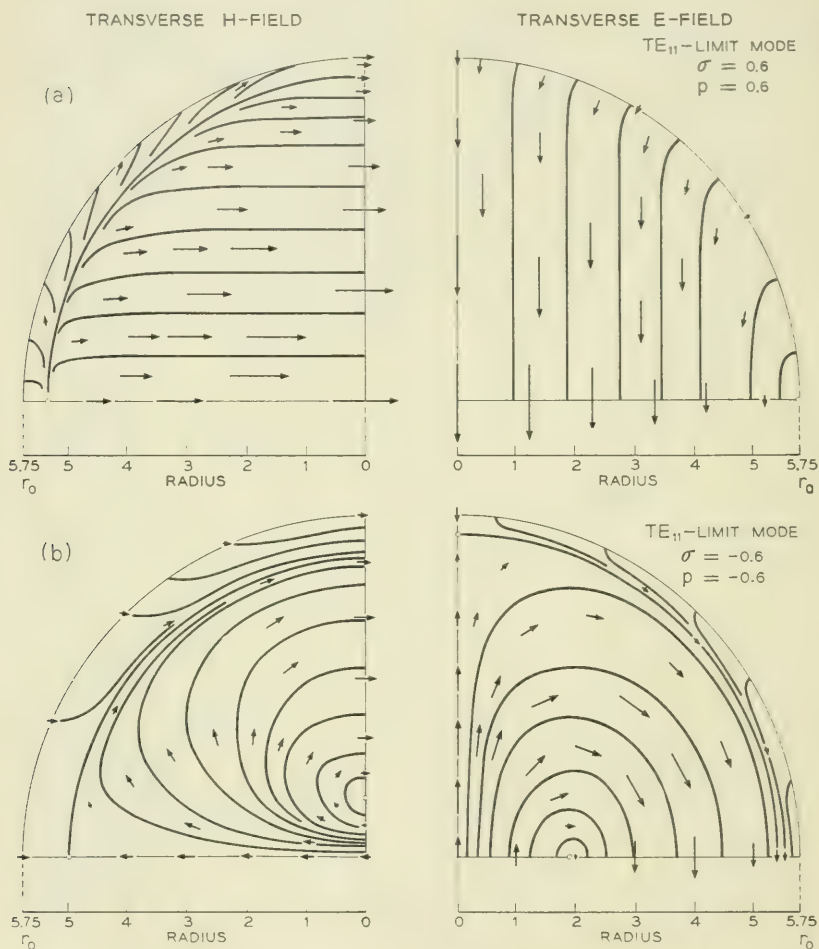


Fig. 10 — Transverse field patterns in a rotating system for four particular cases of Fig. 9. For any pair of E and H patterns the length of the arrows is proportional to field strength, but the patterns for different σ -values have not been given a common normalization. (a), top, TE₁₁-limit mode, $\sigma = 0.6$, $p = 0.6$; (b), bottom, TE₁₁-limit mode, $\sigma = -0.6$, $p = -0.6$; (c), TM₁₁-limit mode, $\sigma = 0.6$, $p = 0.6$; and (d), $\sigma = -0.3$, $p = -0.6$.

they shall fall into an orderly sequence. In Fig. (10) the actual transverse E - and H - field patterns are shown for some representative cases.

It will be noted from Fig. 9 that for both TE- and TM-limit modes, the amplitude patterns for $0 < \sigma < \sigma_0$ resemble those for $\sigma < -1$, while those for $-\sigma_0 < \sigma < 0$ resemble those for $1 < \sigma$. Further, the patterns in the first two ranges of σ are quite similar to the isotropic patterns; those of the latter ranges depart markedly from the isotropic. All of these similarities are more clearly seen for the TE modes than for

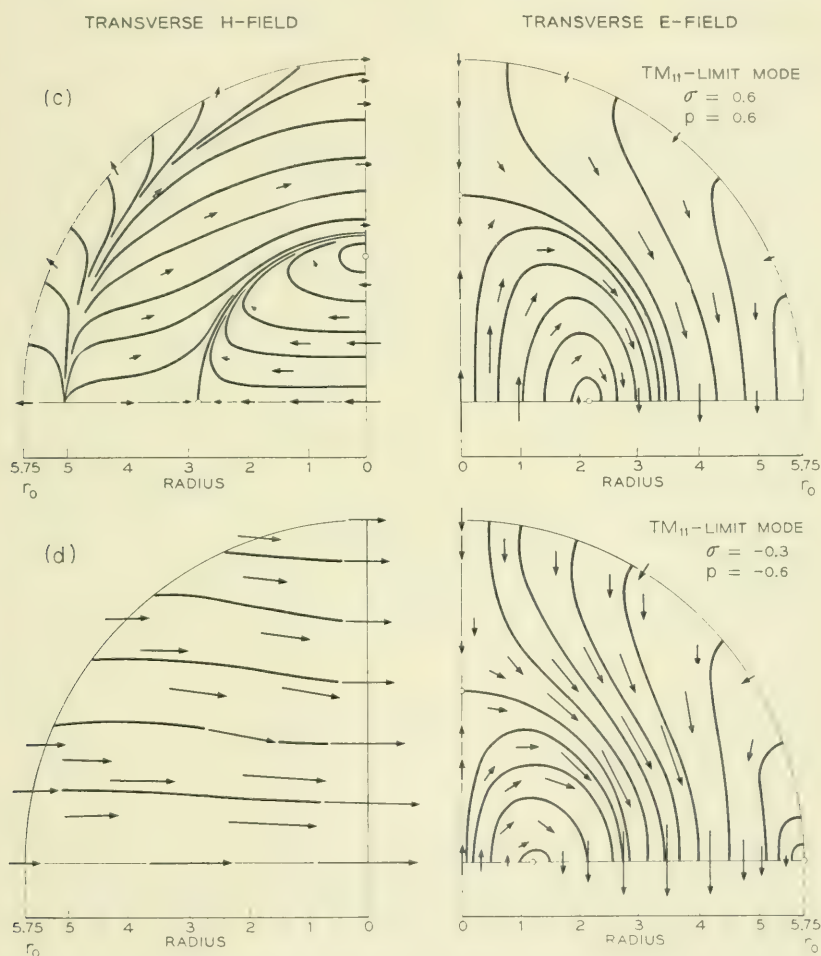


Fig. 10(c), top, and (d), bottom — See Fig. 10.

the TM modes. In Part I where the propagation in guides of very large radius was briefly examined, it may be recalled that in both of the ranges $0 < \sigma < \sigma_0$ and $\sigma < -1$ the field patterns approached those of a plane wave rotating in the same sense as the pattern as a whole. Again, it was found that for $-\sigma_0 < \sigma < 0$ or $\sigma > 1$, the field at most points in the guide was locally rotating in the opposite direction to that of the whole pattern. The points of similarity to the present case are clear and it is also evident that the TE mode more nearly approaches the large guide situation because $r_0 = 5.75$ is much further above the cut-off radius for this mode than it is beyond the cut-off for the TM-mode.

ACKNOWLEDGMENTS

We are indebted to M. T. Weiss for frequent consultations and for permission to use some of the data computed for him. We are also obliged to J. H. Rowen and to S. P. Morgan, Jr., for advice in some matters relating to Part III, and to R. Kompfner for useful discussions concerning the "non-reciprocal helix." R. W. Hamming arranged the integration of the Ricatti equation of Part II.

Special thanks are due to Mrs. A. Rebarber for the many computations relating to Parts I and III, to Mrs. C. A. Lambert for her calculations on Part II, and to Miss M. J. Brannon for a number of results computed for Part III.

REFERENCES

1. A. A. Th. M. Van Trier, Applied Sci. Res., Sec. B, **3**, p. 305, B. Lax, Private communication.
2. R. B. Adler, Res. Lab. of Electronics, M. I. T., Technical Report No. 102, May, 1949.
3. J. H. Rowen, B. S. T. J., **32**, pp. 1333-1369, Nov., 1953.

Bell System Technical Papers Not Published in this Journal

ALLISON, H. W., see MOORE, G. E.

BANGERT, J. T.¹

The Transistor as a Network Element, (Abstract), I.R.E. Trans. ED-1, No. 1, p. 7, Feb. 1954.

BECK, A. C.¹

Microwave Testing with Millimicrosecond Pulses, I.R.E. Trans. MTT-2, No. 1, pp. 93-99, April, 1954.

BICKELHAUPT, C. O.²

Long Lines Signal Communication in the Battle of the Bulge, Telephony, 146, pp. 20-21, 38, May 29, 1954.

BIONDI, F. J.¹

Corrosion Proofing Electronic Apparatus Parts Exposed to Ozone (Abstract), I.R.E. Trans. ED-1, No. 1, p. 65, Feb. 1954.

BOND, W. L.¹

Making Crystal Spheres, Rev. Sci. Instr., 25, No. 4, p. 401, April, 1954.

BRACKETT, G. R.⁴

The Private Branch Exchange, Telephony, 146, pp. 17-18, 49, April 10, 1954.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

⁴ Northwestern Bell Telephone Company.

CHAPIN, D. M.,¹ FULLER, C. S.,¹ AND PEARSON, G. L.¹

A New Silicon p-n Junction Photocell for Converting Solar Radiation into Electric Power, J. Appl. Phys., **25**, No. 5, pp. 676-677, May, 1954.

CLARK, M. A.¹

An Acoustic Lens as a Directional Microphone, I.R.E. Trans. AU-2, No. 1, pp. 5-7, Jan.-Feb. 1954.

CORENZWIT, E., see MATTHAIS, B. T.

CORY, S. I.¹

A New Portable Telegraph Transmission Measuring Set, A.I.E.E. Commun. and Electronics, No. 11, pp. 59-62, March, 1954.

COWAN, F. A.²

Networks for Theater Television, J.S.M.P.T.E., **62**, pp. 306-313, April, 1954.

DEWALD, J. F.¹

Effects of Space Charge on the Rate of Formation of Anode Films, Letter to the Editor, Acta Metallurgica, **2**, pp. 340-341, Mar., 1954.

DITZENBERGER, J. A., see FULLER, C. S.

FEINSTEIN, J.¹

Some Stochastic Problems in Wave Propagation—Part II, I.R.E. Trans. AP-2, No. 2, p. 63, April, 1954.

FLETCHER, R. C.,¹ YAGER, W. A.,¹ PEARSON, G. L.,¹ HOLDEN, A. N.,¹ READ, W. T.,¹ AND MERRITT, F. R.¹

Spin Resonance of Donors in Silicon, Letter to the Editor, Physical Review, **94**, No. 5, pp. 1392-1393, June 1, 1954.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

FULLER, C. S.,¹ STRUTHERS, J. D.,¹ DITZENBERGER, J. A.,¹ AND WOLFSTIRN, K. B.¹

Diffusivity and Solubility of Copper in Germanium, *Phys. Rev.*, **93**, pp. 1182-1189, Mar. 15, 1954.

FULLER, C. S., see CHAPIN, D. M.

GEBALLE, T. H.,¹ AND HULL, G. W.¹

Seebeck Effect in Germanium, *Phys. Rev.*, **94**, No. 5, pp. 1134-1140, June, 1, 1954.

GREEN, E. I.¹

The Decilog — A Unit for Logarithmic Measurements, *Elec. Eng.*, **73**, No. 7, pp. 597-599, July, 1954.

GROSS, A. J., see TANENBAUM, M.

HEFFNER, H.¹

Analysis of the Backward Wave Traveling Wave Tube, *I.R.E. Proc.*, **42**, No. 6, pp. 930-937, June, 1954.

HOHN, F. E.¹

The Conference on Training in Applied Mathematics, *Am. Math. Monthly*, **61**, No. 4, pp. 242-245, April, 1954.

HOLDEN, A. N., see FLETCHER, R. C.

HULL, G. W., see GEBALLE, T. H.

JAFFE, HANS, see MASON, W. P.

LADD, F. E.¹

50 Mc TVI — Its Causes and Cures, Part I, *QST*, **38**, No. 6, p. 21, June, 1954. Part II, *QST*, **38**, No. 7, p. 32, July, 1954.

MASON, W. P., see SHOCKLEY, W.

¹ Bell Telephone Laboratories, Inc.

MASON, W. P.,¹ and JAFFE, Hans⁵

Methods for Measuring Piezoelectric, Elastic and Dielectric Coefficients of Crystals and Ceramics, I.R.E. Proc., **42**, No. 6, pp. 921-930, June, 1954.

MATTHIAS, B. T.,¹ CORENZWIT, E.,¹ and MILLER, C. E.¹

Superconducting Compounds, Letter to the Editor, Phys. Rev., **93**, p. 1415, Mar. 15, 1954.

MATTHIAS, B. T.¹ and CORENZWIT, E.¹

Superconducting Alloys, Letter to the Editor, Phys. Rev., **94**, No. 4, p. 1069, May 15, 1954.

MAY, J. E., JR.¹

Characteristics of Ultrasonic Delay Lines Using Quartz and Barium Titanate, J. Acoust. Soc. Am., **26**, No. 3, pp. 347-355, May, 1954.

MCKAY, K. G.¹

Avalanche Breakdown in Silicon, Phys. Rev., **94**, No. 4, pp. 877-884, May 15, 1954.

MENDEL, J. T.,¹ QUATE, C. F.,¹ AND YOCUM, W. H.¹

Electron Beam Focusing With Periodic Permanent Magnetic Fields. Proc. I.R.E., **42**, pp. 800-810, May, 1954.

MERRITT, F. R., see FLETCHER, R. C.

MILLER, C. E., see MATTHIAS, B. T.

MOORE, G. E.,¹ ALLISON, H. W.,¹ AND WOLFSTIRN, K. B.¹

Reduction of SrO by Methane, J. Chem. Phys, **22**, No. 4, p. 726, April, 1954.

MORIN, F. J.¹

Electrical Properties of NiO, Phys. Rev., **93**, pp. 1199-1204, Mar. 15, 1954.

¹ Bell Telephone Laboratories, Inc.

⁵ Brush Laboratories Company.

MORIN, F. J.¹

Electrical Properties of α -Fe₂O₃, Phys. Rev., 93, pp. 1195-1199, Mar. 15, 1954.

PEARSON, G. L., see CHAPIN, D. M.

PEARSON, G. L., see FLETCHER, R. C.

PFANN, W. G., see TANENBAUM, M.

PRINCE, M. B.¹

Drift Mobilities in Semiconductors: II—Silicon, Phys. Rev., 93, pp. 1204-1206, Mar. 15, 1954.

QUATE, C. F., see MENDEL, J. T.

READ, W. T., see FLETCHER, R. C.

SHOCKEY, W.,¹ AND MASON, W. P.¹

Dissected Amplifiers Using Negative Resistance, Letter to the Editor, J. Appl. Phys., 25, No. 5, p. 677, May, 1954.

SLEPIAN, DAVID¹

Estimation of Signal Parameters in the Presence of Noise, I.R.E. Trans. P.G.I.T.-3, pp. 68-89, Mar. 1954.

STANSEL, F. R.¹

An Improved Method of Measuring the Current Amplification Factor of Junction Transistors, I.R.E. Trans. PG1-3, pp. 41-49, April, 1954.

STRUTHERS, J. D., see FULLER, C. S.

TANENBAUM, M.,¹ GROSS, A. J.,¹ AND PFANN, W. G.¹

Purification of Antimony and Tin by a New Method of Zone Refining, J. Metals, 6, No. 6, pp. 762-763, June, 1954.

THOMAS, D. E.¹

¹ Bell Telephone Laboratories, Inc.

A Point Contact Transistor VHF FM Transmitter, I.R.E. Trans. ED-1, No. 1, pp. 43-52, April, 1954.

TURNER, E. C.⁶

Telephone Growth Forecasting. Part I, Telephony, **146**, No. 24, pp. 17-19, 48, June, 12, 1954. Part II, Telephony, **146**, No. 25, pp. 23-25, June 19, 1954.

VANCE, R. L.¹ AND MAGGS, C.¹

Magnetron Heater Design to Avoid Undesirable Magnetic Effects, (Abstract), I.R.E. Trans. ED-1, No. 1, p. 64, Feb. 1954.

VARNEY, R. N.¹

Liberation of Electrons by Positive-ion Impact on the Cathode of a Pulsed Townsend Discharge Tube, Phys. Rev., **93**, pp. 1156-1160, Mar. 15, 1954.

WALKER, L. R.¹

Stored Energy and Power Flow in Electron Beams, Letter to the Editor, J. Appl. Phys., **25**, No. 5, pp. 615-618, May, 1954.

WICK, R. J.¹

Solution of the Field Problem of the Germanium Gyrator, J. Appl. Phys., **25**, No. 6, pp. 741-750, June, 1954.

WINDELER, A. S.¹

Polyethylene-Insulated Telephone Cable, A.I.E.E. Commun. and Electronics, No. 12, pp. 106-111, May, 1954.

WOLFSTIRN, K. B., see MOORE, G. E.

WOLFSTIRN, K. B., see FULLER, C. S.

YAGER, W. A., see FLETCHER, R. C.

YOCUM, W. H., see MENDEL, J. T.

¹ Bell Telephone Laboratories, Inc.

⁶ New York Telephone Company.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

BENNETT, W.

Telephone-System Applications of Recorded Machine Announcements, Monograph 2213.

BOGERT, B. P.

On the Band Width of Vowel Formants, Monograph 2167.

BROWN, W. L.

N-Type Surface Conductivity on P-Type Germanium, Monograph 2173.

BURTON, J. A., HULL, G. W., MORIN, F. J., and SEVERIENS, J. C.

Effects of Nickel and Copper Impurities on the Recombination of Holes and Electrons in Germanium, Monograph 2193.

BURTON, J. A., PRIM, R. C., SLICHTER, W. P., KOLB, E. D., and STRUTHERS, J. D.

Distribution of Solute in Crystals Grown from the Melt, Monograph 2231.

COY, J. A.

Heat Dissipation from Toll Transmission Equipment, Monograph 2214.

CONWELL, E. M., see DEBYE, P. P.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

DEBYE, P. P.

Electrical Properties of n-type Germanium, Monograph 2220.

ELLIS, W. C., and FAGEANT, JACQUELINE.

Orientation Relationships in Cast Germanium, Monograph 2219

FAGEANT, JACQUELINE, see ELLIS, W. C.

FELCH, E. P., see POTTER, J. L.

Preliminary Development of a Magnetron Current Standard, Monograph 2198.

FELKER, J. H.

Arithmetic Processes for Digital Computers, Monograph 2208.

GOERTZ, M., see WILLIAMS, H. J.

GRISDALE, R. O.

The Properties of Carbon Contacts, Monograph 2206.

HAGSTRUM, H. D.

Electron Ejection from Ta by He^+ , He^{++} , and He_2^+ , Monograph 2148.

HOPKINS, I. L.

The Ferry Reduction and the Activation Energy for Viscous Flow Monograph 2203.

HULL, G. W., see BURTON, J. A.

KARNAUGH, M.

The Map Method for Synthesis of Combinational Logic Circuits, Monograph 2199.

KOLB, E. D., see BURTON, J. A.

LANDER, J. J.

Auger Peaks in the Energy Spectra of Secondary Electrons from Various Materials, Monograph 2215.

LEWIS, W. D.

Electronic Computers and Telephone Switching, Monograph 2187.

LINVILL, J. G.

A New RC Filter Employing Active Elements, Monograph 2221.

MAY, J. E.

Characteristics of Ultrasonic Delay Lines Using Quartz and Barium Titanate Ceramic Transducers, Monograph 2223.

McSKIMIN, H. J.

Measurement of Elastic Constants at Low Temperatures by Means of Ultrasonic Waves Data for Silicon and Germanium Single Crystals and for Fused Silica, Monograph 2171.

MORIN, F. J., see BURTON, J. A.

PENNELL, E. S.

A Temperature-Controlled Ultrasonic Solid Delay Line, Monograph 2222.

PFANN, W. G.

Change in Ingot Shape During Zone Melting, Monograph 2218.

PIERCE, J. R.

Spatially Alternating Magnetic Fields for Focusing Low-Voltage Electron Beams, Monograph 2169.

POTTER, J. L., see FELCH, E. P.

PRIM, R. C., see BURTON, J. A.

PRINCE, M. P.

Experimental Confirmation and Relation Between Pulse Drift Mobility and Charge Carrier Drift Mobility in Germanium, Monograph 2168.

READ, W. T., JR.

Dislocations and Plastic Deformations, Monograph 2216.

REISS, H.

Chemical Effects due to the Ionization of Impurities in Semiconductors, Monography 2172.

RYDER, E. J.

Mobility of Holes and Electrons in High Electric Fields, Monograph 2159.

SCHNETTLER, F. J., see WILLIAMS, J. H.

SEVERIENS, J. C., see BURTON, J. A.

SHERWOOD, R. C., see WILLIAMS, J. H.

SHOCKLEY, W.

Transistor Physics, Monograph 2217.

SLICHTER, W. P., see BURTON, J. A.

SNOKE, L. R.

Soil-Block Bioassay of a Creosote Containing Pentachlorophenol Monograph 2212.

STRUTHERS, J. D., see BURTON, J. A.

STRUTHERS, J. D., see THURMOND, C. D.

THURMOND, C. D.

Equilibrium Thermochemistry of Solid and Liquid Alloys of Germanium and of Silicon, Monograph 2189.

TIEN, P. K.

Traveling-Wave Tube Helix Impedance, Monograph 2209.

WILLARD, G. W.

Ultrasonically Induced Cavitation in Water — A Step-by-Step Process, Monograph 2170.

WILLIAMS, H. J., SHERWOOD, R. C., GOERTZ, M., AND SCHNETTLER, F. J.

Stressed Ferrites having Rectangular Hysteresis Loops, Monograph 2200.

Contributors to this Issue

ORSON L. ANDERSON, B.S., M.S. and Ph.D., University of Utah, 1948, 1949 and 1951; Institute of Rate Processes, University of Utah, 1949-1952; Bell Telephone Laboratories, 1952-. Dr. Anderson has been engaged in the investigation of mechanical and electrical properties of solids, with emphasis on glasses, and in studies of the mechanism of plastic flow of amorphous bodies. A member of the mechanics division of the Mathematics Department, he is now studying the strength and flow properties of glasses under high pressure. Member of American Physical Society, American Ceramic Society and Society of Glass Technology.

J. K. GALT, A.B., Reed College, 1941; Ph.D., M.I.T., 1947; O.S.R.D., M.I.T. and Harvard University, 1943-1945; National Research Council Fellow, Bristol, England, 1947-1948; Bell Telephone Laboratories, 1948-. Dr. Galt has been engaged in research on the properties of solids, especially of ferrites, with emphasis on their magnetic properties. Fellow of the American Physical Society and member of Phi Beta Kappa.

W. P. MASON, B.S. in E.E., University of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921-. Dr. Mason has been engaged principally in investigating the properties and applications of piezoelectric crystals, in the study of ultrasonics, and in mechanics. Fellow of the American Physical Society, Acoustical Society of America and Institute of Radio Engineers and member of Sigma Xi and Tau Beta Pi.

J. L. MERRILL, JR., B.S. and M.S., Pennsylvania State University, 1928 and 1930; Elliot Research Fellow, 1928-1930; American Telephone and Telegraph Company, 1930-1934; Bell Telephone Laboratories, 1934-. Mr. Merrill spent his first years with the Laboratories on transmission features of such projects as the time and weather announcement systems and operator training programs. During World War II, he engaged in planning system operation of air raid warnings as well as work on tactical wire and radio networks for the armed forces. Since the war he has been concerned with the design and application of negative

impedance repeaters for the improvement of exchange transmission. He holds several patents and is the author of numerous technical articles. Member of Theta Alpha Phi.

IRAD S. RAFUSE, B.S. in E.E., Cooper Union, 1927; Columbia University; Western Electric Company, 1920-1925; Bell Telephone Laboratories, 1925-. He worked on the development of high quality vertical disc recording for a number of years before turning to measurements and testing of switching apparatus. During World War II he engaged in the development of sonar equipment in cooperation with the N.D.R.C. and the Bureau of Ships and Ordnance from which he received a commendation for his work. He was later in charge of a group engaged in the development of a new wire-spring, multi-contact relay and now is in charge of a group developing glass sealed switches.

ARTHUR F. ROSE, B.S. in E.E., Colorado College, 1914; American Telephone and Telegraph Company, 1914-. Mr. Rose immediately joined the General Engineering Department of the A.T.&T. Co. Upon completing the student training course for new employees, he was assigned to the development work then under way on the New York-San Francisco route which culminated in the first transcontinental telephone service in 1915. As a result of this initial acquaintance with telephone repeaters, he continued in transmission work dealing particularly with these devices. In 1919, when the General Engineering Department was divided and the Operating and Engineering Department formed, Mr. Rose was assigned to the group that was concerned primarily with the application of repeaters and carrier systems in toll engineering. In 1939 he was transferred to the Plant Extension Section and in 1953 returned to the Transmission Section as Exchange Transmission Engineer.

J. O. SMETHURST, B.S. in Communications, Tufts College, 1929; Bell Laboratories, 1929-. For many years he was concerned with overseas telephony, concentrating especially on control terminals for radio telephone circuits. During World War II he was associated with various government projects and after the war he worked on NIKE. Since 1953 he has concentrated on E2 and E3 repeaters.

HARRY SUHL, B.Sc., University of Wales, 1943; Ph.D., Oriel College, University of Oxford, 1948. Admiralty Signal Establishment, 1943-46; Bell Telephone Laboratories, 1948-. Dr. Suhl conducted research on the properties of germanium until 1950 when he became concerned with

electron dynamics and solid state physics research. His current work is in the applied physics of solids. Member of the American Institute of Physics and Fellow of the American Physical Society.

LAURENCE R. WALKER, B.Sc. and Ph.D., McGill University, 1935 and 1939; University of California, 1939-41. Radiation Laboratory, Massachusetts Institute of Technology, 1941-1945; Bell Telephone Laboratories, 1945-. Dr. Walker has been primarily engaged in research on microwave oscillators and amplifiers. At present he is a member of the physical research group concerned with the applied physics of solids. Fellow of the American Physical Society.

H E B E L L S Y S T E M

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII

NOVEMBER 1954

NUMBER 6

Waveguide as a Communication Medium S. E. MILLER 1209
A Governor for Telephone Dials — Principles of Design

W. PFERD 1267

In-Band Single-Frequency Signaling
A. WEAVER AND N. A. NEWELL 1309

Centralized Automatic Message Accounting System G. V. KING 1331

The Wave Picture of Microwave Tubes J. R. PIERCE 1343

Theory of Open-Contact Performance of Twin Contacts
M. M. ATALLA AND MISS R. E. COX 1373

Bell System Technical Papers Not Published in this Journal 1387

Recent Bell System Monographs 1392

Contributors to this Issue 1398

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

S. BRACKEN, *Chairman of the Board,*
Western Electric Company

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. McNEELY, *Vice President, American Telephone*
and Telegraph Company

EDITORIAL COMMITTEE

W. H. DOHERTY, *Chairman*

F. R. LACK

A. J. BUSCH

W. H. NUNN

G. D. EDWARDS

H. I. ROMNES

J. B. FISK

H. V. SCHMIDT

E. I. GREEN

G. N. THAYER

R. K. HONAMAN

J. R. WILSON

EDITORIAL STAFF

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXIII

NOVEMBER 1954

NUMBER 6

Copyright, 1954, American Telephone and Telegraph Company

Waveguide as a Communication Medium

By S. E. MILLER

(Manuscript received March 23, 1954)

The circular electric wave in round metallic tubing has an attenuation coefficient which decreases as the frequency of operation is increased. A corollary to this behavior is the fact that any preselected attenuation coefficient can in theory be obtained in any predetermined diameter of pipe through the choice of a suitably high carrier frequency. The attenuation which is characteristic of microwave radio repeater links, about 2 db/mile, is in theory attainable in a copper pipe of about 2" diameter using a carrier frequency near 50,000 mc.

Scale-model transmission experiments, conducted at 9,000 mc, showed average transmission losses about 50 per cent above the theoretical value. These extra losses were due to (1) roughness of the copper surface and (2) transfer of power from the low-loss mode to other modes which can also propagate in the pipe.

The latter effect may have serious consequences on signal fidelity because power will transfer (at successive waveguide imperfections) from the signal mode to unused modes and, after a time delay, back to the signal mode. This effect has been studied experimentally and theoretically, and it is concluded that (1) either mode filters must be inserted periodically to absorb the power in the unused modes of propagation, or (2) the medium itself must be modified to continuously provide large attenuations for the unused modes of propagation. The latter approach is attractive in that it also provides a solution to the problem of bending this form of low-loss guide.

The general outlook, based on present knowledge, is that a waveguide system might transmit baseband widths as large as 100 to 500 mc using a rugged modulation method such as PCM. Some form of regeneration is likely to be required at each of the repeaters, which may be spaced on the order of 25 miles. A total rf bandwidth of about 40,000 mc may be available in a single guide.

TABLE OF CONTENTS

Introduction.....	1210
Ordinary Versus Circular Electric Waves.....	1211
Theoretical Characteristics of the Circular-Electric Wave.....	1213
Some Results of Transmission Experiments.....	1219
Mode Conversion and Reconversion as a Signal-Loss Effect.....	1229
Mode Conversion and Reconversion as an Interference Effect.....	1230
Analysis for Continuous Mode Conversion.....	1239
Direct Evaluation of Mode Conversion Magnitudes.....	1243
The Bend Problem.....	1247
Improved Forms of Circular Electric Waveguide.....	1250
Surface Roughness.....	1252
Circular Electric Wave and Millimeter Wave Techniques.....	1253
Modulation Methods.....	1256
Conclusion.....	1259
Appendix — Theoretical Analysis for Continuous Mode Conversion.....	1261

INTRODUCTION

The circular electric wave in round metallic tubing possesses a property so unique that some early research workers doubted the reality of the wave. This unique property is an attenuation coefficient which, in a given pipe, *decreases* without limit as the frequency of operation is *increased*. In parallel wire, coaxial, or ordinary waveguide lines the "skin effect" at the surface of the conductor causes the loss to increase as the frequency increases indefinitely, so the predicted circular-electric-wave loss characteristic aroused considerable interest as soon as it was discovered by S. A. Schelkunoff and G. C. Southworth in the early 1930's. Since that time considerable work has been done at Holmdel to explore the reality of the circular electric wave and to evaluate its usefulness to the Bell System. It is the purpose of this paper to report on the status of this work and to give a description of some of the basic characteristics of circular electric wave propagation.

The Bell System is interested in knowing whether waveguide can be used as a long distance communication medium in the manner in which coaxial cable or the radio relay system is now employed. Our interest in long distance waveguides is due in part to the fact that radio-wave propagation through the atmosphere becomes progressively more severely handicapped by rain, water vapor and oxygen absorptions at

frequencies above 12,000 mc. Use of the spectrum above the 10,000-20,000 mc region seems to require a sheltered transmission medium.

Circular electric wave transmission may also find application in short connecting links, such as between subscribers requiring very broad band circuits, between two central offices as a multi-channel carrier link, or between a radio relay antenna site and a somewhat remote transmitter-receiver location chosen for accessibility.

In each of these cases, the broad bands available in the microwave portion of the spectrum, the complete shielding afforded by waveguides generally, combined with the low-loss properties of the circular electric wave would seem to provide an ideal transmission medium. We therefore seek knowledge of the precision required in the waveguide and some indication of general system complexity to facilitate a judgment as to whether the cost will be competitive.

ORDINARY VERSUS CIRCULAR ELECTRIC WAVES

Let us approach a discussion of circular electric waves by considering their relation to the waveguides which are now used in our radio relay systems and which found widespread use in the radars of World War II. The vast majority of waveguides in commercial use now are rectangular in cross section and have dimensions large enough so that one and only one wave-type, usually called the "dominant mode", can propagate. To simplify this discussion, such waveguides will be called ordinary waveguides. Ordinary waveguides are analogous to coaxial or parallel-wire lines in many respects. Because only one mode can propagate, departures from an absolutely straight tube of constant cross section show up as reactance effects only. A dent in the side wall of the guide or of the coaxial, an abrupt change in cross section, or a twist or bend of the line all appear as non-dissipative reflection effects which may be cancelled at one frequency (or in one band of frequencies) by the addition of another compensating reactance at a point suitably located. A great many of the components used in ordinary waveguides, including the frequency selective filters, depend on such reactance cancellation effects in order to achieve satisfactory operation.

Since the techniques for employing ordinary waveguides have been thoroughly explored, it is natural to inquire as to whether we can use them for communication purposes. We do use ordinary waveguides in lengths of the order of 100 feet and more to connect the antennas and repeaters in the 4,000 mc (TD-2) radio relay system. The attenuation is excessive, however, for long-distance applications. The particular type

of brass rectangular waveguide used for TD-2 transmission lines has attenuation in excess of 50 db per mile, and use of the very best copper would only reduce the theoretical loss to about 40 db per mile. In order to reduce the loss in ordinary waveguides, just as in coaxial or parallel wire lines, one must go to lower frequencies. In particular, the theoretical loss at a carrier frequency near 1000 mc is about 2 db per mile, which is about the same as the transmitter-to-receiver attenuation in our radio relay systems. The waveguide in the 500–1000 mc region would have cross-sectional dimensions on the order of one foot, would be cumbersome to handle and would involve rather large material cost. In addition, it turns out that such a waveguide would be useful in signal bandwidths only a few mc wide as a result of delay distortion, which will be discussed further in the ensuing discussion. Thus, we have concluded that ordinary waveguide is not very attractive as a transmission medium over distances on the order of a mile or more.

It is true that the attenuation in any hollow metallic waveguide can be reduced to any desired extent at a given frequency by making the cross sectional area larger by a suitable factor. The penalty is that the transmission medium becomes capable of propagating energy in several characteristic ways, known as modes. The striking feature of a multi-mode transmission medium is that energy in one mode is entirely independent and unaltered by the presence or absence of energy in one of the other modes. This situation is sketched diagrammatically in Fig. 1. Energy can theoretically propagate between 1 and 1', between 2 and 2', and between 3 and 3' at the *same time* and in the *same frequency band* without mutual interference. The separate modes represent independent transmission lines which occupy the same space. The distinguishing features of the various modes in a multi-mode waveguide are: (1) Velocity of propagation or phase constant, (2) Attenuation coefficient, and (3) Configuration of electric and magnetic field lines within the waveguide.

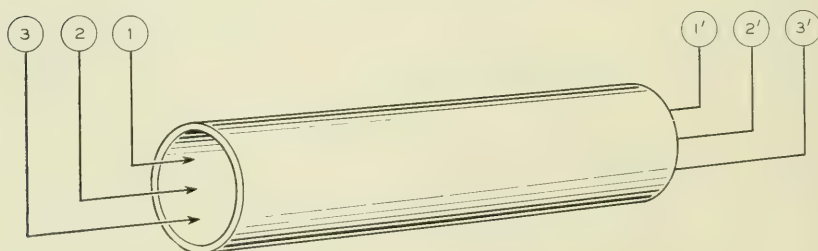


Fig. 1 — Diagram of multi-mode waveguide transmission.

The fact that it is necessary to use a waveguide whose dimensions are large enough to permit the existence of a number of modes has far-reaching influence on the research being discussed here. Practically, the independence between the various modes of propagation is limited by tolerances of various kinds. In the multimode waveguide, changes in cross section or bends or twists require design attention with regard to mode purity as well as with regard to impedance match, and it is not permissible to insert arbitrarily shaped probes or irises for impedance matching purposes as is the common practice in ordinary waveguides. This means that a complete new technique is required for the old components, such as frequency-selective filters, hybrids, and attenuators, as well as for a new series of components such as pure mode generators and mode filters.

THEORETICAL CHARACTERISTICS OF THE CIRCULAR ELECTRIC WAVE

Since it has been found necessary to use a waveguide in the multimode region in order to get the desired losses in a reasonable size waveguide, we may inquire as to which of the modes is best suited to our problem. At a given frequency the loss for any one of the modes may be reduced as much as is desired by making the cross sectional area of the guide large enough, but there is a mode for which the loss decreases with increasing guide size much more rapidly than for any other mode. This is the circular electric (TE_{01}) mode in straight round pipe. It turns out that no current flows in the direction of propagation in the metallic walls of a straight round pipe carrying the circular electric mode. It is the absence of current in the direction of propagation which permits the circular-electric-wave attenuation to decrease indefinitely as the frequency increases, and this difference between ordinary transmission lines and the circular-electric wave is further illustrated in Fig. 2. In the familiar parallel-wire line the electric field extends directly from one conductor to the other, resulting in charge accumulations at half-wave intervals along the axis of propagation and associated conduction currents in the copper wires. These conduction currents in the direction of propagation do not diminish as the frequency of operation increases, since they are associated with the energy transmitted to the end of the transmission line. With the circular electric wave the electric field lines close upon themselves, are always tangential to the conducting wall, and do not result in a charge accumulation on the walls due to the main energy flow. The wall currents which do flow are merely sufficient to prevent the propagating energy from spreading out as it would if the metallic walls

were not present, but these currents decrease rapidly with increasing frequency in a given waveguide size. Only in a straight circular pipe can all of the electric field lines close upon themselves, and only in the straight circular pipe does the attenuation approach zero at infinite frequency.*

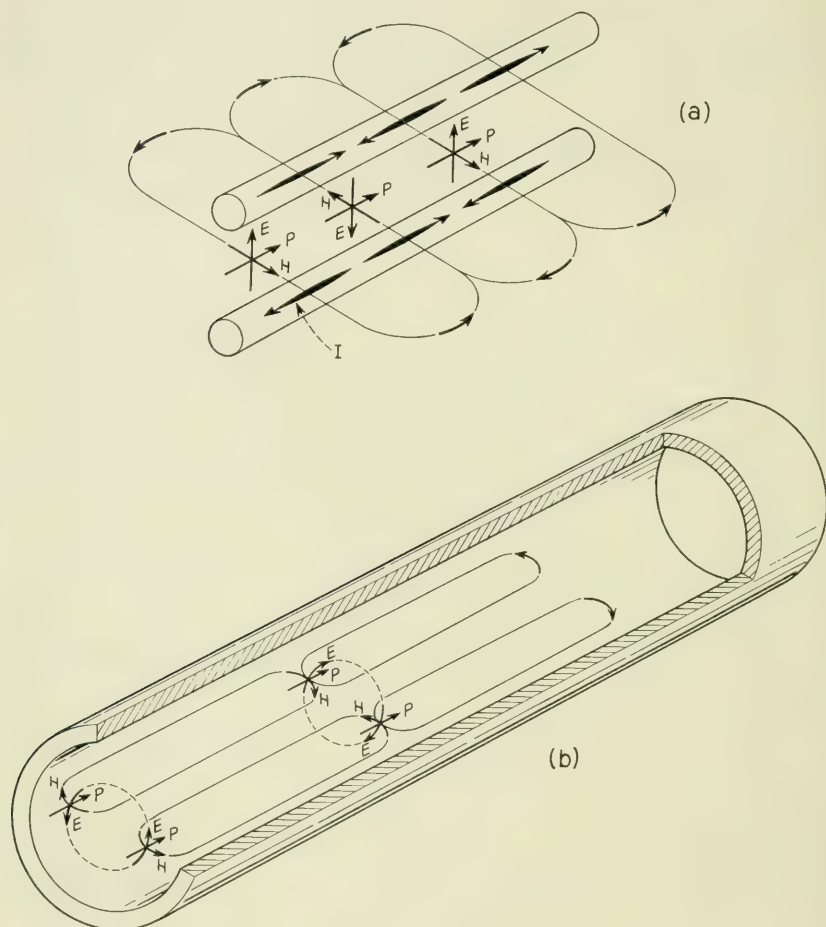


Fig. 2 — Sketch of the magnetic intensity (H), electric intensity (E) and Poynting Vector (P) for parallel-wire and circular-waveguides. Because the main energy flow (P) in the circular electric waveguide is associated with electric field lines that close upon themselves and do not produce accumulations of charge on the metal walls, the wall currents and associated losses are very small.

* For further discussion, see G. C. Southworth, *Principles and Applications of Waveguide Transmission*, D. Van Nostrand, Inc. pp. 175-178.

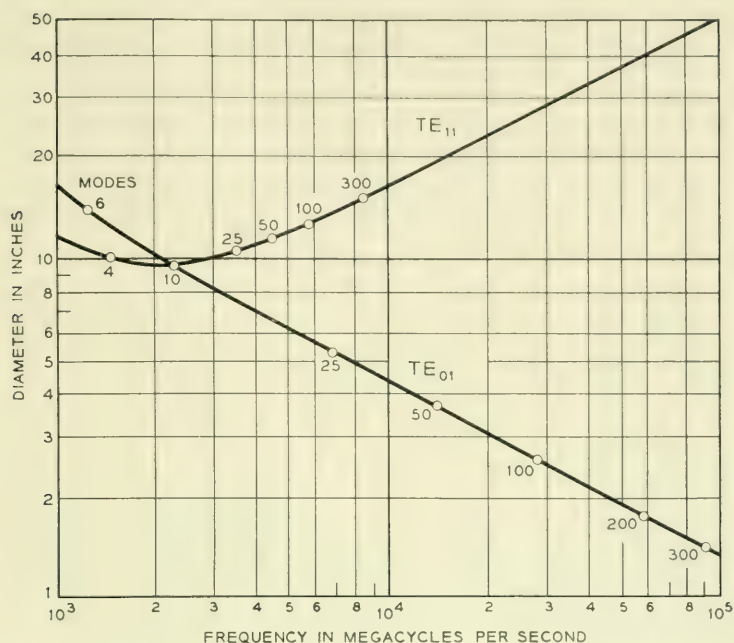


Fig. 3 — Round guide diameter versus frequency for attenuation of 2 db/mile.

As a consequence of the unusual loss versus frequency characteristic of the circular electric wave, the diameter required in order to achieve a given loss decreases as the carrier frequency increases. This is illustrated by the curves labeled TE_{01} in Figs. 3 and 4. All other waveguide modes (except higher-order circular-electric waves, TE_{0m}) have a characteristic of the general form sketched for the dominant wave (TE_{11}) also shown in Figs. 3 and 4. The longitudinal wall currents contribute a loss component which rises at increasing frequencies due to skin effect; this accounts for the positive slope of the TE_{11} curve at the right-hand side of Figs. 3 and 4. The negative slope of the TE_{01} curves and of the left-hand portion of the TE_{11} curves in Figs. 3 and 4 is a consequence of losses associated with the wall currents which prevent the wave from spreading as it would in an unbounded medium; these currents and the losses associated with them decrease as the operating frequency becomes farther removed from cut-off.

For a loss of 2 db per mile Fig. 3 shows that a waveguide 7" in diameter is required in the frequency band near 4,000 mc where the TD-2 system operates. Whereas this may not be prohibitive in a connecting

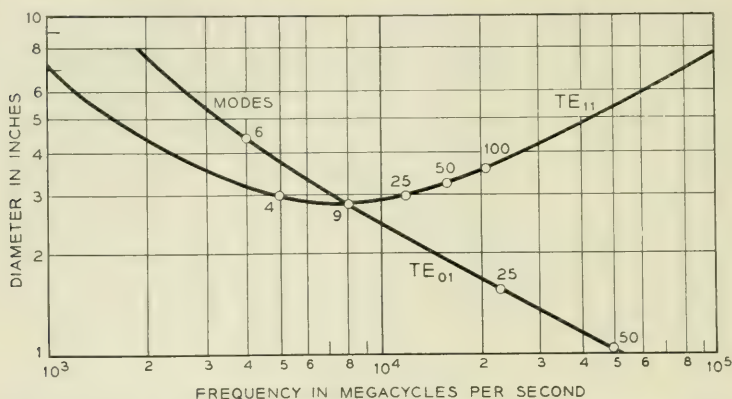


Fig. 4 — Round guide diameter versus frequency for attenuation of 13.2 db/mile (0.25 db/100 ft.).

link application, the waveguide size is definitely too large for long-distance application. In the vicinity of 50,000 mc, however, Fig. 3 shows that the required waveguide size is on the order of 2", and this is comparable to the size of the present standard 8-pipe coaxial cable. From these simple calculations, it is evident that carrier frequencies in the vicinity of 50,000 mc or more are very desirable for long-distance waveguide applications in order to minimize the size of the waveguide.

Other reasons for wanting a high carrier frequency arise from a consideration of bandwidth. Any hollow conductor waveguide has a cutoff characteristic of the form sketched in Fig. 5. Above cutoff the group velocity approaches asymptotically to the velocity in an unbounded medium composed of the dielectric used in the waveguide. Because the group velocity varies across the frequency band, a signal being transmitted in a waveguide will experience delay distortion; the components transmitted at $f_0 \pm \Delta f$ (Fig. 5) would be delayed compared with their relation at the input to the line. When this delay is 180° a baseband signal of frequency Δf would be severely distorted regardless of the modulation method, and this condition may be regarded as an upper limit to the usable bandwidths in the waveguide unless correction for delay distortion is employed. This particular type of phase distortion has been analyzed in unpublished papers, first by D. H. Ring and later by S. Darlington. The work of Darlington leads to the following relation between the parameters of the waveguide and the baseband width f_B associated with the 180° phase difference noted above:

$$f_B = \frac{304 f^{1/2} (1 - v^2)^{3/4}}{v L^{1/2}} \text{ (cps)}$$

where f is the carrier frequency (cps), v is the ratio of the waveguide cutoff frequency to the carrier frequency, and L is the line length in miles.

From the above relation, the maximum baseband width available has been calculated for one mile of line as a function of carrier frequency with the loss held constant at 2 db per mile and 13.2 db per mile, and the results are plotted in Fig. 6. The conditions for these curves are directly comparable to those for which Figs. 3 and 4 were calculated. At a carrier frequency near 50,000 mc, the circular electric wave in 2-inch diameter pipe makes available a baseband width on the order of 500 mc for one mile of line, or 100 mc for 25 miles of line. At lower frequencies, with the waveguide enlarged to hold the loss constant, there is less bandwidth available.

The 13.2 db per mile condition (in a smaller diameter of waveguide) permits the use of approximately one-half the bandwidth available at the 2 db per mile condition.

The above design considerations are the basis for concluding that the most attractive communication possibility is the use of carrier frequencies on the order of 50,000 megacycles and associated waveguide

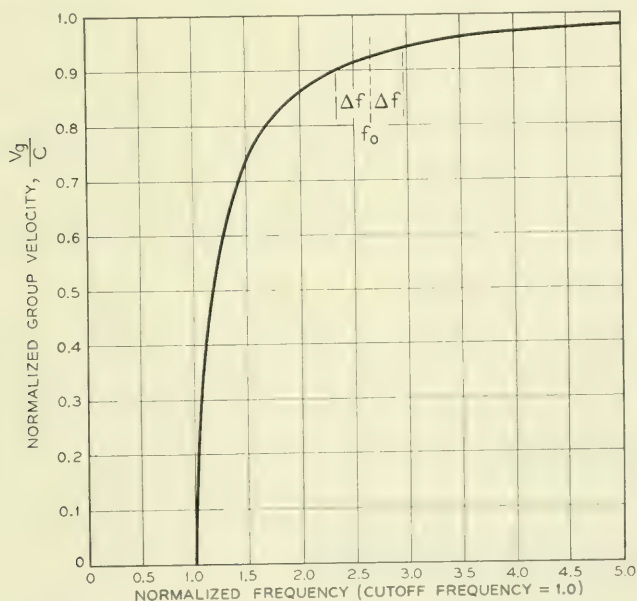


Fig. 5 — Normalized group velocity versus normalized frequency for hollow metallic waveguides.

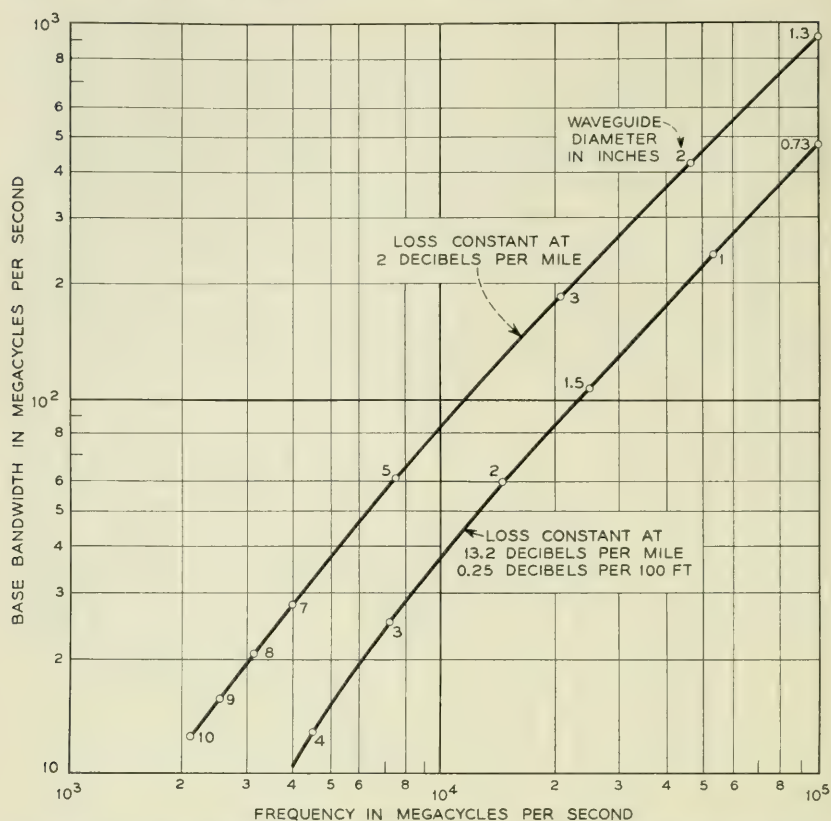


Fig. 6 — Base band width per channel versus frequency (for one-mile of waveguide) using the circular electric mode, loss held constant by varying pipe diameter.

diameters on the order of 1 to 2 inches. Having selected the general region which appears attractive, it is important to know the communication characteristics of a given waveguide. Figs. 7 and 8 show the attenuation and bandwidth characteristics respectively for a 2-inch round copper tube. The theoretical loss (Fig. 7) for the TE_{01} wave varies from 4 db per mile to 0.64 db per mile over the frequency range from 30,000 to 100,000 mc. The higher order circular electric waves have somewhat more attenuation but may be found useful as auxiliary transmission channels for short distances.

It is interesting and important, as will be pointed out later, to note that several of the higher order transverse electric waves have attenuation constants which may be within a factor of 3 to 6 times the attenuation for the circular electric wave. These waves, TE_{12} and TE_{13} , have

very little wall current flowing in the direction of propagation and therefore approximate the low-loss properties of the circular electric wave family. Under certain conditions of mode coupling, which will be described at a later point, it is undesirable for the medium to be able to propagate modes with attenuation coefficients comparable to that of the mode which is used for communication purposes.

SOME RESULTS OF TRANSMISSION EXPERIMENTS

Transmission experiments have been conducted on the 500-ft waveguide line shown in Fig. 9.* Supports for the line were set in concrete

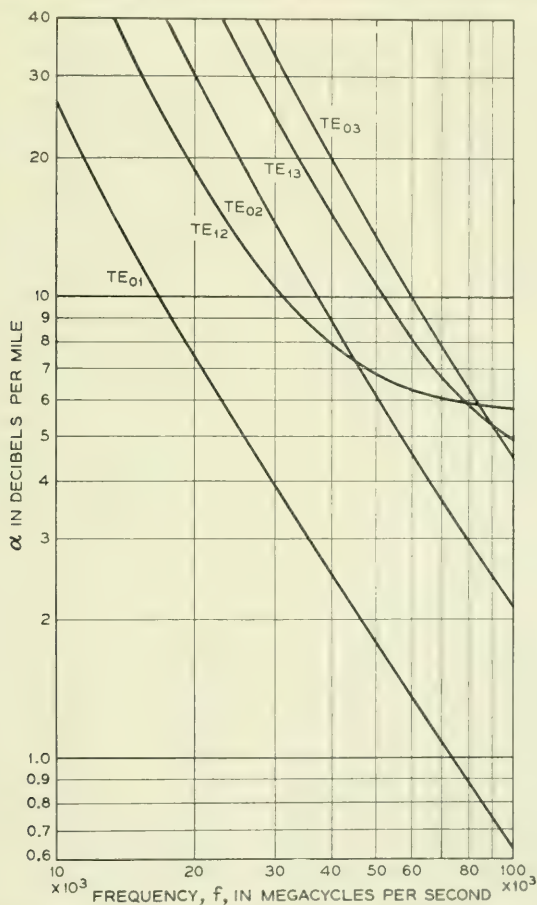


Fig. 7 — Attenuation versus frequency for a 2" diameter round waveguide.

* This is the same line used for the work reported in Reference 1. Some of the experiments described in Reference 1 are also described here in order to furnish background for the new material.

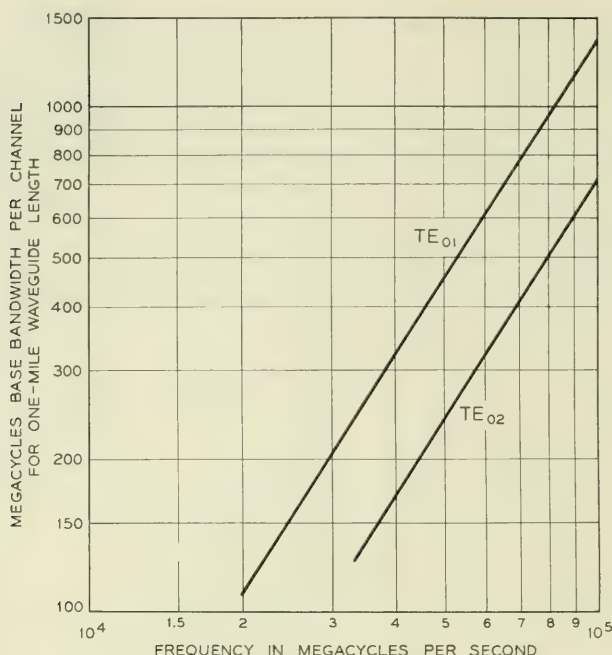


Fig. 8 — Base band width per channel versus frequency for a 2" diameter pipe (one-mile waveguide length).

and optically aligned so as to provide a waveguide straight within about $\frac{1}{8}$ " over its entire length. The philosophy behind this installation was the familiar one of providing for experimental purposes, as close to the ideal line as possible so that deviations could be created in a controlled manner. The inside diameter is about 4.73", chosen to obtain the desired theoretical loss of about 2 db per mile at 9,000 mc, where measuring equipment was readily available. The difference between the major and minor inside diameters of the pipe was in the range 0.005" to 0.008" at the ends of the sections which averaged 20 ft in length. At the time this work was initiated, in 1946, generators of higher frequencies which would permit the use of smaller waveguides had not yet become available for use in this research.

Tests were conducted on this line using a technique due to A. C. Beck,¹ and involving the layout of equipment shown in Fig. 10. Short bursts of *RF* energy approximately $\frac{1}{10}$ microsecond in duration, were injected into the line at intervals of about three hundred microseconds. Except for two small holes through which to couple to the transmitter and

receiver, the waveguide line was short-circuited at both ends. The injected 10^{-10} microsecond pulse occupied at any instant a space interval of 100 feet and therefore this pulse, while travelling from one end to the other between the short circuits, produced at the receiver coupling hole spurts of energy corresponding to the time when the pulse passed the sending end. Each such pulse passing through the receiver coupling hole was amplified, detected, and placed as a vertical deflection on the oscilloscope. The horizontal deflection on the oscilloscope was a linear time

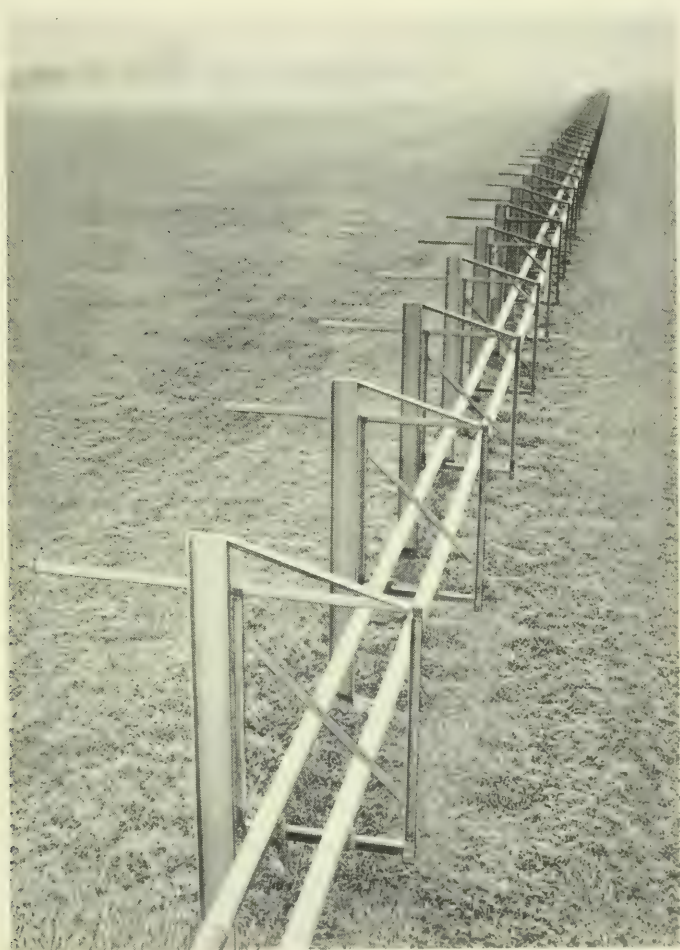


Fig. 9 — Experimental waveguide installation, 5" diameter Holmdel line.

base having a duration of a few microseconds. In order to observe the received pulse after a selected number of trips back and forth down the waveguide, a variable delay was placed between the trigger and the oscilloscope horizontal deflection. The discussion which follows will refer to photographs of the oscilloscope under different conditions. By way of preparation, it may be stated that the pulse transmitted through the small hole in the end plate of the waveguide will excite a large number of modes. There are, at 9,000 mc, approximately 40 modes which can propagate in this waveguide. Also, the coupling through the holes is so weak and the energy lost due to dissipation in the shorting plates is so small as to represent an attenuation which is small compared with the theoretical wall loss in the 500-foot long line. Therefore, as the pulse shuttles back and forth in the line, it will decay as though it had travelled on a straight long section of waveguide made up of 500-foot long segments identical to the single 500-foot section actually constructed.

Fig. 11 shows a photograph of the oscilloscope displaying the time interval immediately following the transmitted pulse. The pulse at the extreme left represents the transmitted pulse which passes directly from the transmitter hole to the receiver hole on the end plate of the waveguide. The blank time interval immediately following the transmitted pulse is about one microsecond long and represents the time of travel of energy down to the far end of the 500-foot line and back to the sending end. During this interval no pulses were received because the joints in the line produce little reflection. The first pulse after the transmitted pulse represents energy travelling in the mode which has the highest

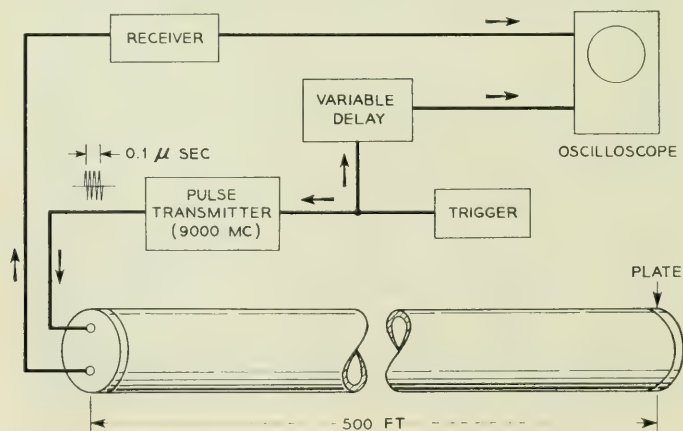


Fig. 10 — Diagram of equipment used for pulse tests of waveguide transmission.

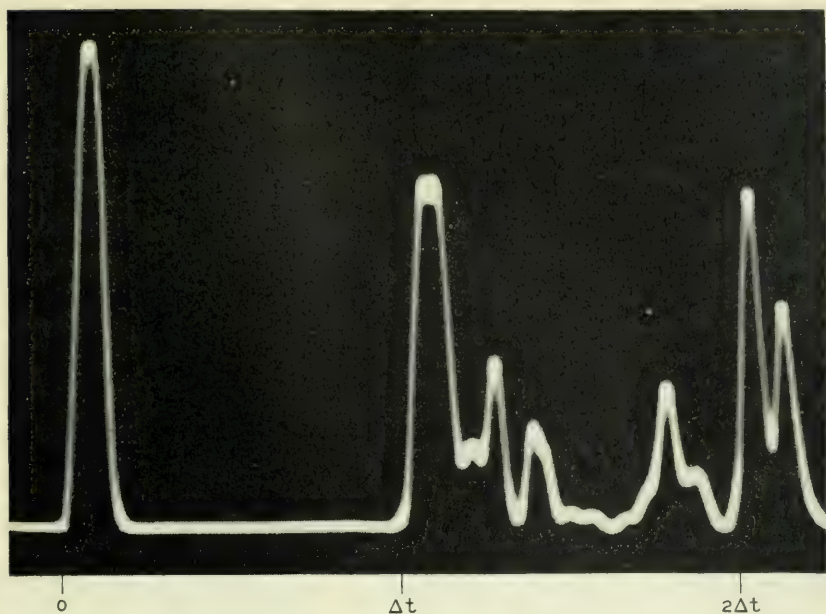


Fig. 11 — Photograph of cathode-ray tube presentation during the time interval immediately following the transmitted pulse.

group velocity. Pulses immediately following this first received pulse represent energy travelling in other modes whose velocities are lower and which therefore require more time for the one round trip of travel. At the time $2\Delta t$, we begin to observe pulses which have made two round trips in the line. If the transmitted pulse width were short enough, we could theoretically identify the mode in which the energy travelled by observing the time of arrival, since the velocities of propagation and the distance are known parameters. The $\frac{1}{10}$ microsecond pulse used in these experiments is not short enough to allow this kind of resolution on an individual mode basis. Something on the order of five or six modes have velocities so nearly the same that they cannot be resolved as separate pulses with the $\frac{1}{10}$ microsecond pulse after a single round trip in the 500-foot line.

Fig. 12 represents the same condition as Fig. 11, except that the horizontal time base has been changed to display the interval 0 to $14\Delta t$ instead of the interval 0 to $2\Delta t$. Fig. 12 shows fewer pulses in the time interval $6\Delta t$ to $12\Delta t$ than in the time interval 0 to $6\Delta t$. This is because energy travelling in some modes is attenuated more rapidly than that in other modes. For time delays greater than $10\Delta t$ the received pulses

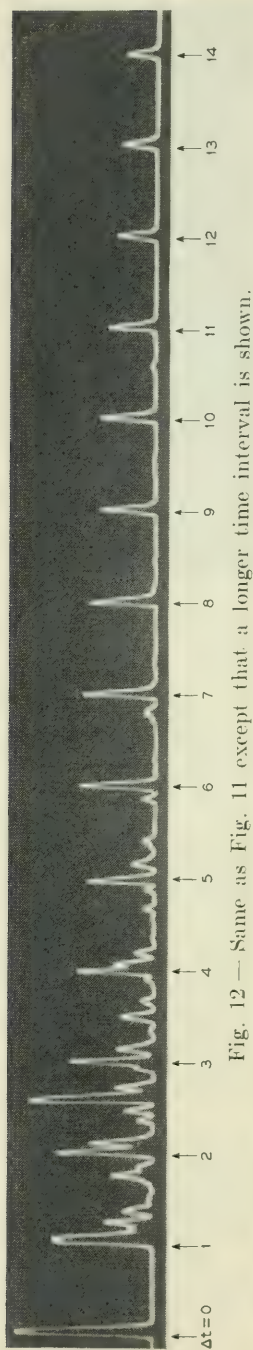


Fig. 12 — Same as Fig. 11 except that a longer time interval is shown.

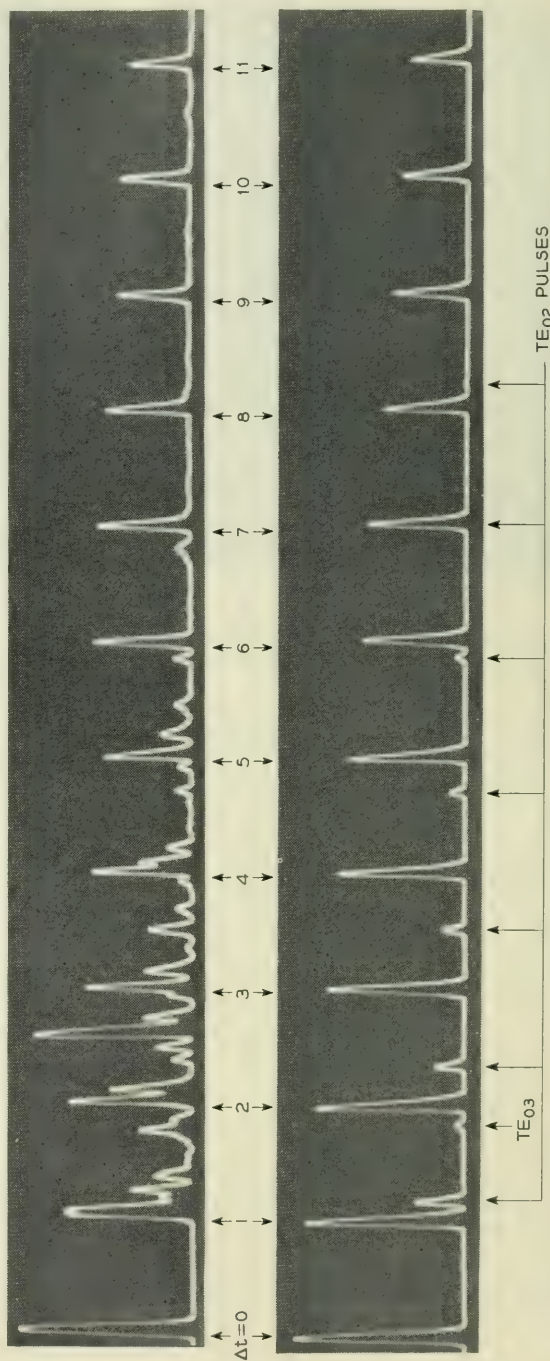


Fig. 14 — The effect of a mode filter in the 500-foot line. The unlabeled pulses in the lower part are in the TE_{01} mode.

appear at regular intervals and with smoothly decaying amplitude. This behavior indicates that the major portion of the energy in the line was travelling in a single mode, and we deduced that this mode was TE_{01} as follows: We observed that the velocity of propagation was near that for the TE_{01} mode by measuring the absolute time between pulses, averaged over many round trips. This excluded all but about 6 modes whose cut-off frequencies are near that of TE_{01} . Measurement of transmission loss was made by observing the rate of decay of the received pulses averaged over 10 or more round trips. The measured loss was found to be approximately 3 db per mile compared to a theoretical value of 1.9 db per mile for TE_{01} propagation. It follows that propagation must have

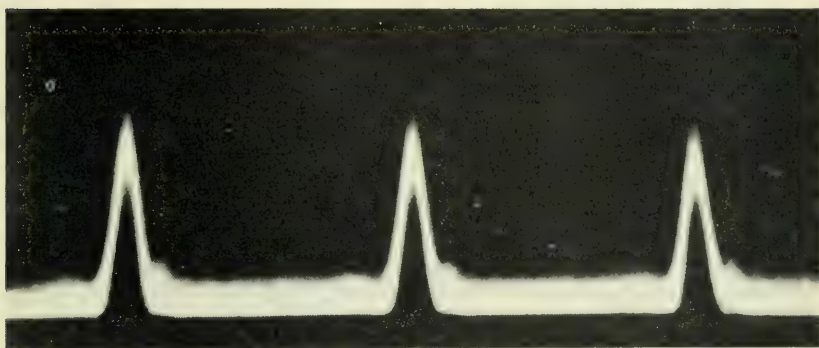


Fig. 13 — Record of pulses after 40 miles of repeated traversal over the 500-foot line.

been taking place in the TE_{01} mode, for all other modes near TE_{01} in velocity have theoretical losses well in excess of the observed value.

To summarize the effects shown in Fig. 12, a great many modes including TE_{01} were launched by exciting the waveguide through a small aperture in the end plate. All these modes propagated back and forth in the line for a while, but due to the fact that TE_{01} has appreciably less loss than the other modes, the energy remaining in the line after a suitable time delay was substantially all in the TE_{01} mode. This permitted measuring TE_{01} loss over a distance of many miles by allowing the energy to traverse the 500-foot line many times.

Fig. 13 records three successive trips of a pulse which had travelled up and down the 500-foot waveguide for a total distance of 40 miles. The pulse shape was still essentially the same as that of the transmitted pulse, although background noise had become clearly visible. We cer-

tainly can conclude from this that circular electric wave transmission over great distances is possible.

The long waveguide line also provided a very convenient way of demonstrating additional multi-mode transmission effects. For example, in a multimode medium we may use mode filters. One such filter may have a very low loss for the circular electric waves but very high loss to other waves. Such mode filters have been built, and Fig. 14 shows the transmission changes which result when introducing one. The upper half of Fig. 14 shows the photograph of the oscilloscope in the time interval 0 to $11\Delta t$ with no mode filter in the line. (This is the same as Fig. 12.) On introducing the mode filter into the waveguide, the received signal is altered as shown in the lower half of Fig. 14. We observed that energy propagating in the undesired modes has largely disappeared; it was absorbed by the mode filter.

There are still a few small pulses in the lower half of Fig. 14 which cannot be in the TE_{01} mode because of their time position. Starting at time $1.15\Delta t$, there is a series of regularly spaced pulses in Fig. 14 labeled TE_{02} , and a single small pulse labeled TE_{03} . The geometric placement of resistive material in the mode filter leads us to anticipate low filter losses for the entire circular electric (TE_{0n}) family of modes, and therefore the extra pulses were suspected of being in higher-order circular electric modes. Only two such modes, TE_{02} and TE_{03} , were above cut-off. The TE_{03} pulse was tentatively identified by noting that its group velocity was 55 to 60 per cent of that of the TE_{01} pulses. High attenuation in the TE_{03} mode prevented additional TE_{03} pulses from being observed.

In the case of the TE_{02} series of pulses, it was possible to get a fairly accurate measure of relative group velocity, confirming the identification as TE_{02} . Note that the seventh TE_{01} pulse coincides with the sixth TE_{02} pulse, and that the pulse at $7\Delta t$ shows on the TE_{01} train as being too large in amplitude.

The smooth decay of the TE_{01} train in the lower half of Fig. 14 in the interval Δt to $6\Delta t$ is in marked contrast to the corresponding pulse train in the upper half of the figure, and is graphic illustration of the important effects that mode filtering can produce.

Another very important transmission observation appeared during Mr. Beck's experiments with the 5" diameter line. He observed that the attenuation, as measured by the amplitude decay of the shuttling pulse, was a function of the position of the piston at the far end of the line. Translations of the far end piston on the order of 10 to 40 centimeters changed the overall transmission from a condition in which the original

pulse shape was preserved for as many as 40 miles of travel (Fig. 13), to a condition wherein the shape of the pulse was badly distorted after only 3 or 4 miles of travel. This general behavior is illustrated by the photographs shown in Fig. 15. We will concentrate for the moment on the top two rows of photographs which record the pulse transmission in the bare waveguide as a function of distance of pulse travel for both favorable and unfavorable piston settings. However, all of the rows of the photographs were taken under such conditions as to permit direct comparison.

The photographs at the extreme left end represent the outgoing pulse and the first echo pulses which travelled one round trip in the line, approximately 340 yards. All the other photographs show two principal pulses which record two successive trips of the pulse as it passed the transmitting end (Fig. 10). The second photograph from the left represents the pulse as it passed the sending end after 10 and 11 round trips. The third picture from the left records the 20th and 21st trip, the fourth picture the 30th and 31st trip, etc. The numbers placed directly beneath the individual photographs represent the relative sensitivity of the receiver for that particular photograph. Reference receiver sensitivity was taken as the condition under which the 10th and 11th trip in the bare waveguide were recorded with a favorable piston setting (0 db beneath the photograph), and the designation -6 db under the adjacent photograph indicates that 6 db more receiver sensitivity was used in the latter case. The relation between display amplitude and actual pulse amplitude was approximately square law. The distance of pulse travel associated with each of the pictures is given at the bottom of the figure.

Comparing the top two rows representing favorable and unfavorable piston positions, we note that the attenuation was appreciably different — the values being 2.6 db per mile and 3.1 db per mile respectively. Serious distortion of the transmitted pulse also occurred for the unfavorable piston setting. Since the receiver in the experiment was sensitive to very many modes, one might suspect that the spurious pulses which appeared at more than 7,000 yards with the unfavorable piston setting might represent energy present in some of the other modes. Actually, all of the pulses and wiggles shown in the photographs at ranges greater than 3,500 yards were in the circular electric mode. This was deduced by first noting that every two successive pulses were not appreciably different from each other (see Fig. 15). If some of the distortion effects shown by the received pulse were due to energy being received in modes other than the signal mode, then successive pulses would be different in shape because the various modes have different phase constants. The very gradual change in pulse shape which did

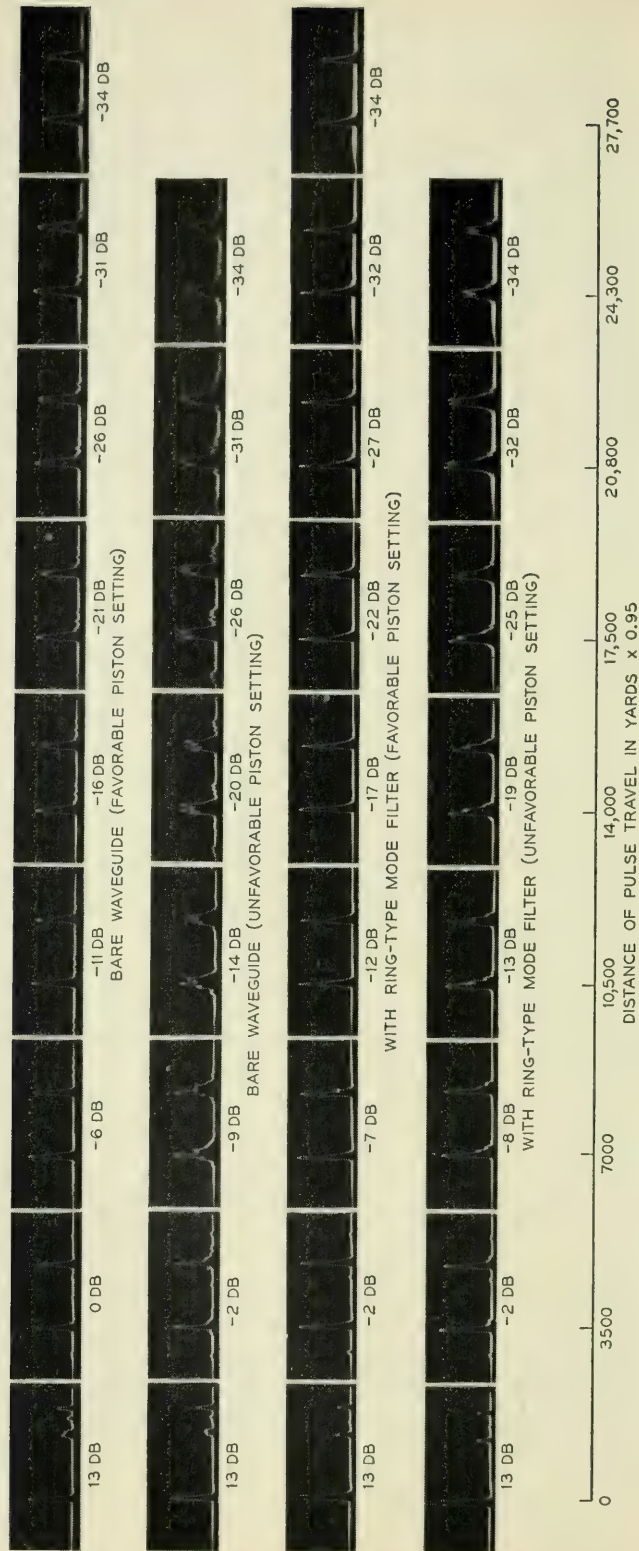


Fig. 15 — Photographs of pulse transmission with both favorable and unfavorable piston settings at the end of the 500-ft. waveguide.

occur as the pulse travelled up and down the line had the general form of an amplitude component which led or lagged the signal pulse by a constant interval but which gradually increased in amplitude with increasing distance that the signal pulse had travelled. Note, for example, the growth of spurious peaks before and after the signal pulse in row 2 of Fig. 15 at 3,500, 7,000, 10,500 and 14,000 yards of travel.

The explanation of this behavior involves transfer of energy from the circular electric mode to one or more of the unused modes of propagation and reconversion of the same energy back to the circular electric mode. This process is one of the characteristic features of multimode waveguide systems and is discussed at greater length in the following sections.

MODE CONVERSION AND RECONVERSION AS A SIGNAL-LOSS EFFECT

In beginning discussion of the mode conversion-reconversion phenomena, we take an idealized case of a dissipationless waveguide containing two deformities, Fig. 16. We assume a $c-w$ signal entirely in the circular electric mode entering this waveguide. After passing the first deformity there will be energy present in some other mode, and this is designated as TX_1 . When the combination of $TE_{01} + TX_1$ strikes the second deformity, another conversion takes place and the output will be a large TE_{01} component, two smaller components in the unused mode, TX_1 and TX_2 , and a still smaller circular electric wave component, TE_{01}' , which is due to reconversion of energy from TX_1 to the circular electric wave in traversing the second deformity. It can be shown² that for the proper distance between two identical symmetrical deformities, the wave emerging from the line may be purely circular electric; the two components TX_1 and TX_2 cancel each other under this condition. Another separation between the deformities results in a maximum energy transfer from circular electric to other modes. Therefore, it follows that any mechanism which varies the effective spacing between conversion points will produce conversion loss variations. This accounts for the change in the attenuation of the circular electric wave pulses in Fig. 15 as a function of the far end piston setting.

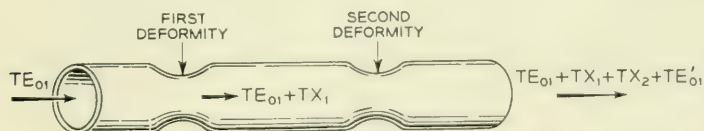


Fig. 16 — A distorted waveguide and the associated mode-conversion signal-loss effects.

MODE CONVERSION AND RECONVERSION AS AN INTERFERENCE EFFECT

The mode conversion-reconversion phenomena can also produce a signal distortion or interference effect. Fig. 17 shows another idealized waveguide containing two deformities, but in this case we assume a *short pulse* in the circular electric wave as the input signal. The amplitude-time plots below the waveguide show the energy present in the circular electric and unused waves respectively at various physical points along the line. The key item in this diagram is the time displacement between the converted energy TX and the circular electric wave energy at the input to the second deformity. This time difference appears as a result of propagation over an identical line length at two different group velocities which, in general, the circular electric and unused modes will possess. Since the signal and unused mode pulses strike the second deformity at different times the second conversion process results in energy appearing back in the circular electric wave at a time separated from the signal pulse itself. When the distance between deformities is too short for the pulses to be resolved at the second deformation, the result will be a distortion of the signal pulse rather than the appearance of a separate pulse.

The above very much simplified picture of the mode conversion and reconversion effects allows one to visualize several general properties of this phenomenon:

1. In general, there will be a large number of unused modes which will be coupled to the signal mode through the various imperfections in the transmission line. Since these unused modes have unequal phase constants, and since the imperfections will be randomly spaced along the line, the reconverted signal pulses in a time-division system will be spread

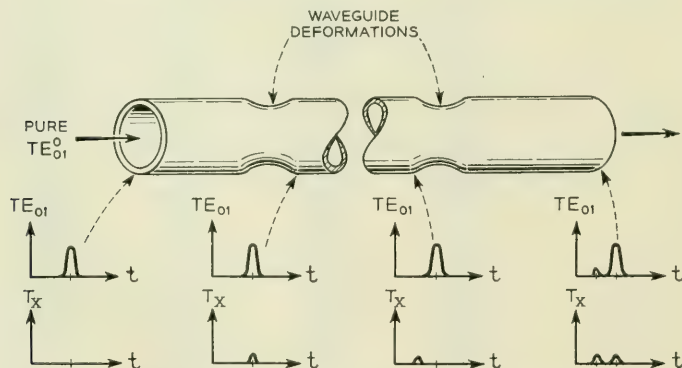


Fig. 17 — Signal interference effects due to mode conversion and reconversion.

out on the time scale rather than appear as a single well-defined distortion pulse.

2. Because some of the unused modes of propagation have group velocities greater than the group velocity of the circular electric wave, the reconverted energy pulses may reach the receiving end of the transmission system before the signal pulse itself appears.

3. The reconverted energy will, in general, be out of phase with the signal from which it was derived. When the differential time of travel in the unused mode is short, the principal effect of the conversion-reconversion process will be to distort the signal wave. When the differential time of travel in the unused mode becomes as large as the reciprocal of the modulation frequencies involved, the reconverted energy will appear more like an echo. In a time-division system such echo pulses would interfere with the pulses representing other signal components. Because of the large number of conversions contributing at random time delays, this "echo-interference" may be unintelligible. In this sense the interference may be thought of as a noise effect, just as multi-channel cross-talk due to amplitude non-linearities in a single sideband AM system may be thought of as noise.

4. The general case of a signal pulse, both preceded and succeeded by a series of reconverted energy pulses, is sketched in Fig. 18. It is quite apparent that if the reconverted signal pulses are allowed to become of the same order as the signal pulse itself, even a pulse code modulation system will be rendered inoperative. Other types of modulation will experience difficulty at appreciably smaller magnitudes of reconverted energy.

5. The level of the reconverted energy relative to the signal is determined by the transmission medium. It is not possible to avoid this interference by using more power at the sending end of the transmission link, for the interference rises with the signal. The need for low-noise receivers is just as acute as in other transmission systems, because better noise figure means that correspondingly less power is required from the transmitter.

6. If the loss to the mode TX is very large in the region between successive waveguide imperfections, the TX pulse can be attenuated to a negligibly small value before reaching the second deformation, thus preventing any significant reversion back to the signal mode.

7. Limits can be placed on the time delay between the signal energy and the reconverted energy returned to the signal mode. The lower limit on this time delay is obviously zero, corresponding to a series of imperfections very close together. The upper limit can be taken as the differ

ence between the times it takes the signal mode and the unused mode to travel a distance in which there is a large difference between the ohmic losses in the unused-mode and signal mode. This concept may be expanded as follows: Energy is transferred at an imperfection located at coordinate z_1 on the axis of propagation, producing an amplitude in unused mode x . At z_2 on the axis of propagation the amplitude in mode x will be attenuated relative to the signal-mode amplitude at the same point by the factor

$$e^{-(\alpha_x - \alpha_1)(z_2 - z_1)}$$

where α_x and α_1 are the normal heat loss coefficients in mode x and the signal mode respectively. When the exponential factor is small enough (i.e., z_2 large enough), reconversion will no longer be important compared to reconversion near z_1 . For order of magnitude we might assume that 10 db more attenuation for the x -mode amplitude than for the signal-mode amplitude would render further reconversion unimportant.

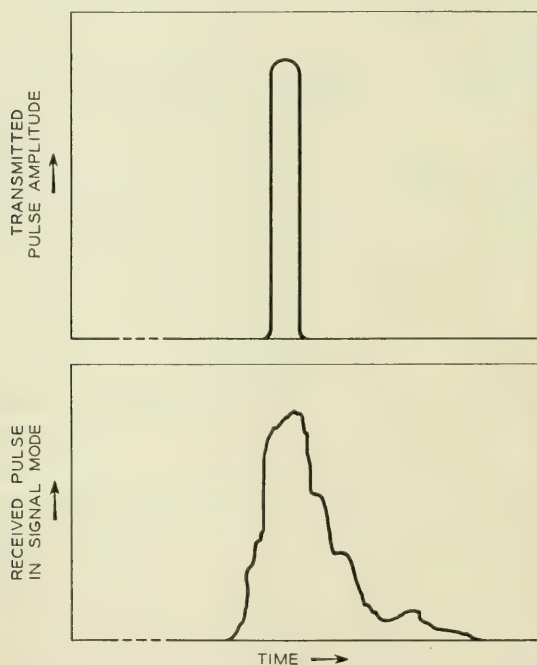


Fig. 18 — Schematic of signal distortion due to conversion and reconversion effects in a line with randomly placed conversion points.

Then we know the distance $(z_2 - z_1)$ from

$$(z_2 - z_1) = \frac{1.15}{(\alpha_x - \alpha_1)}$$

The corresponding "upper limit" on time delay between the signal and the reconverted energy in the signal mode is

$$t = (z_2 - z_1) \left(\frac{1}{v_x} - \frac{1}{v_s} \right) \quad (1)$$

where v_x and v_s are the group velocities in the mode x and the signal mode.

It is well known that the unused modes of a circular electric waveguide have attenuation coefficients which are appreciably larger than attenuation coefficients for the circular electric wave itself. In the light of this attenuation to the unused modes plus the fact that the reconverted energy has undergone two mode conversion losses before it reaches the signal mode again, one might wonder whether the mode conversion-reconversion phenomenon would really be an important effect. The first indication that the magnitudes of the reconversion amplitudes are significant came during experimental work on the 5" diameter 9,000-mc line, described above in connection with Fig. 15. A more quantitative theoretical discussion which follows shows that the reconversion phenomena will continue to be important even when mode filters are introduced into the line.

The effects of the mode conversion-reconversion process are very similar to the effects of multipath transmission through the atmosphere. In microwave radio there are under unusual fading conditions 2 or 3 subsidiary signals, and these are representative of propagation over different path lengths in space but at the same velocity of propagation. In the waveguide, there will in general be a large number of subsidiary transmission paths, each of which corresponds to the identical distance of propagation but at velocities of propagation which are different for the various modes. The radio multipath phenomenon exists only occasionally, whereas the waveguide multipath phenomenon is a steady characteristic present 100 per cent of the time. Long-distance radio transmission in the 6-20 mc region by way of the ionosphere encounters multipath effects more like those expected in the waveguide, with the exception that waveguide multi-path effects are expected to have far greater short-time stability.

Quantitative relations describing the conversion-reconversion process may be derived by considering an infinitely long waveguide composed of

a series of identical sections each containing a single mode-conversion discontinuity at the midpoint. (The actual 500-foot line contains many conversion points, as will be discussed later, and is effectively repeated over and over as the pulse traverses the identical line many times.) A very short signal-pulse of unit amplitude is assumed as an input to the idealized line containing identical conversion points. After the first conversion point the amplitude in the unused mode is k and the amplitude in the signal mode is $(1 - k^2)^{1/2}$, where k is a measure of the size of the conversion irregularity. There is no reconverted wave at this point since there is no input to the first conversion point in the unused mode. After the second conversion point, however, there is a reconverted-wave amplitude $k^2 \epsilon^{\theta_x}$, where θ_x is defined below. After the n^{th} conversion point, it may be shown that the amplitude in the signal mode is

$$\epsilon^{(n-1)\theta_1} (1 - k^2)^{n/2} \quad (2)$$

the amplitude in the unused mode is

$$\begin{aligned} & k(1 - k^2)^{\frac{(n-1)}{2}} \epsilon^{(n-1)\theta_x} \\ & + k(1 - k^2)^{\frac{(n-1)}{2}} \epsilon^{(n-2)\theta_x + \theta_1} \\ & + k(1 - k^2)^{\frac{(n-1)}{2}} \epsilon^{(n-3)\theta_x + 2\theta_1} \\ & + \text{etc. to } + k(1 - k^2)^{\frac{(n-1)}{2}} \epsilon^{(n-1)\theta_1} \end{aligned} \quad (3)$$

and the reconverted wave amplitude is

$$\begin{aligned} & (n-1)k^2(1 - k^2)^{\frac{(n-2)}{2}} \epsilon^{\theta_x + (n-2)\theta_1} \\ & + (n-2)k^2(1 - k^2)^{\frac{(n-2)}{2}} \epsilon^{2\theta_x + (n-3)\theta_1} \\ & + (n-3)k^2(1 - k^2)^{\frac{(n-2)}{2}} \epsilon^{3\theta_x + (n-4)\theta_1} \\ & + \text{etc. to } k^2(1 - k^2)^{\frac{(n-2)}{2}} \epsilon^{(n-1)\theta_x} \end{aligned} \quad (4)$$

in which

z = distance between adjacent conversion points

α_1 and α_x are the heat loss coefficients (applying to wave amplitudes) for the signal and unused modes respectively.

β_1 and β_x are the phase constants for the signal and unused modes respectively.

$\theta_1 = -(\alpha_1 + j\beta_1)z$

$$\theta_x = -(\alpha_x + j\beta_x)z$$

k = amplitude of the converted wave for a single conversion point and for unit wave incident on the conversion point.

In deriving the series represented by (2), (3) and (4), it is assumed that all of the converted power travels in the forward direction, and that reflection effects are negligible. These conditions are typical of imperfections in practical multi-mode waveguides.

When the input pulse for the idealized line is sufficiently short, the various terms of (3) and (4) (representing successive conversions and reconversions) are non-overlapping pulses. It is instructive to write down the ratio of the signal-wave amplitude to the reconverted wave amplitude which is separated from the amplitude of the signal wave at the same point by the time difference $z(1/v_x - 1/v_1)$, in which v_x and v_1 are group velocities. This ratio is [ratio of (2) to the first term of (4)]

$$\frac{(1 - k^2)}{(n - 1)k^2 \epsilon^{-(\alpha_x - \alpha_1)z}} \quad (5)$$

It is clear from this ratio that the signal wave may be smaller than the reconverted wave if n is sufficiently large. Physically, what happens is that the reconverted amplitude created at each successive conversion point adds in phase with the reconverted wave amplitude present at that point due to previous conversions and reconversions. This happens of course because the line contains identically spaced conversion points. With random location of conversion points, a less severe build-up of reconverted wave energy would certainly occur.

The first and second reconverted pulses [i.e., the first and second terms in (4)] are separated by the time difference $z(1/v_x - 1/v_1)$ and the ratio of the amplitude of the second to the first reconverted pulse is

$$\frac{(n - 2)}{(n - 1)} \epsilon^{-(\alpha_x - \alpha_1)z} \quad (6)$$

The largest amplitude existing in the unused mode at a point immediately following the n^{th} conversion is the component converted at the n^{th} conversion point and the ratio of this component to the amplitude of the signal pulse after the n^{th} conversion is:

$$\frac{k}{(1 - k^2)^{1/2}} \quad (7)$$

Note that this ratio is independent of n . The unused-mode amplitude converted at the first conversion point is, after n trips and relative to

the signal pulse at the same point,

$$\frac{k}{(1 - k^2)^{1/2}} \epsilon^{-(n-1)(\alpha_x - \alpha_1)z} \quad (8)$$

These expressions show that the largest unused mode amplitude at any point in the line will be the one most recently converted from the signal wave and will be nearest to the signal pulse in time position.

The sketch shown in Fig. 19 represents schematically the signal pulse, the unused mode pulse and the reconverted pulse amplitudes after n trips past the conversion point. It is interesting to note that the most recent (the n^{th}) conversion to the unused mode appears in a time position close to the signal pulse whereas the most recent reconversion appears at a time far removed from the signal pulse.

Let us investigate the ratios (5) through (8) under conditions representative of those in the 5" diameter waveguide line. Row 2 of Figure 15 shows that after 40 trips down and back on the 500-foot line, i.e., after 80 trips past the center of this line (where there is assumed a single conversion point), the amplitude of the reconverted pulses which appear just before and just after the signal pulse are about equal to the signal

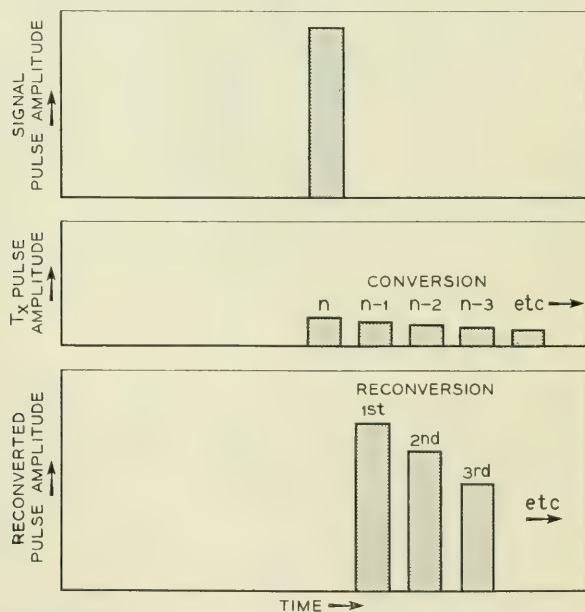


Fig. 19 — Sketch of pulse amplitudes in a line having equally spaced conversion points.

pulse. We may then set the ratio of signal-wave amplitude to the reconverted-wave amplitude, as given by (5), equal to unity at $n = 80$. We know the theoretical values of the heat loss coefficients α_r and α_1 for the unused and signal modes respectively. Allowing modest increase for surface roughness effects, we will assume the heat loss for the signal (TE_{01}) wave to be 0.2 db between conversion points (about 500 feet apart) and the loss for the unused mode to be about five times this or 1 db between conversion points. Substituting these numbers into (5) we can solve for the conversion coefficient k which is necessary to obtain the observed equality between reversion and signal wave amplitudes at the $n = 80$ point. Such a calculation yields a value of k of 0.117 and this implies a conversion loss which is 32 per cent of the heat loss for the signal wave. Direct observations of mode conversion (to be described) show that the conversion losses in the 5" diameter line must be at least as large as this. Therefore, we have confirmed that the basic mechanism which we are discussing can indeed account for a break-up of the signal pulse of the general type actually observed.

The calculated ratio of the second reconverted pulse compared to the first reconverted pulse from (6) for the $n = 80$ condition described above is only -0.5 db, which shows that if we really had a single conversion point which could add exactly in-phase in the manner assumed, we would have in our photographs an even worse series of reconverted pulses than actually do appear.

Again for the $n = 80$ condition in the 5" line the calculated amplitude of the unused mode component relative to the signal component at the same point is -18.5 db for the n^{th} conversion and, with the square law response of the display system, it is to be expected that such an amplitude would not be observed. This fits satisfactorily with the observation that the pulses in the 5" line, after 40 round trips, are all in the circular electric mode even though the conversion-reconversion process is significant.

As already implied, the mode conversion situation for the actual 500-foot line, which is represented by the photographs of Fig. 15, is more complicated than the idealized line just analyzed. First, note that in the case of the nonoverlapping pulses in the idealized line the position of the far end piston (i.e., the spacing between the conversion points) has no critical effect on the conversion and reversion amplitudes. In the experiment (Fig. 15), far-end piston movements of a foot or so caused very significant changes in the observed conversion and reversion. This is a consequence of the fact that in the physical 500-foot line, there were a large number of conversion points, and the pulse width

was not small enough to produce non-overlapping effects. If all of the conversions in the 500-foot length were between TE_{01} and a single additional mode, then it may be shown that the piston position which would cause in-phase addition of components from one conversion point on successive traversals of the 500-foot line would also cause in-phase addition for the components from all of the other conversion points. However, in the physical line significant conversion takes place between TE_{01} and three or more other modes, each of which has a different phase constant and each of which requires a different piston setting for in-phase (or out-of-phase) addition of components generated on successive traversals. Whereas it is not possible to relate the shuttle-pulse observations of Fig. 15 to a simple quantitative theory, the general behavior of the experimental line is in qualitative agreement with the conversion-reconversion explanation.

The favorable piston setting in Fig. 15 represents an infinite line in which approximate amplitude cancellation of mode conversion effects is achieved, whereas the unfavorable piston setting corresponds to an infinite line in which approximate amplitude addition of mode conversion effects is experienced. Since the former is optimistic and the latter is pessimistic compared to what might be expected in a physical long line, it has been our practice to average the loss data obtained at the favorable and unfavorable piston positions.

The expression (5) for the ratio of the signal pulse to the reconversion pulse shows that appreciable loss between the conversion points, represented by the factor $e^{-(\alpha_r - \alpha_1)z}$, is an effective way of reducing the reconverted wave effects. During the early experiments it was found that mode filters did reduce the influence of the far end piston. In Fig. 15, rows 3 and 4, show records of the 5" line pulse transmission with a single mode filter at the sending end. The mode filter did not completely eliminate the phasing effects of the piston, and this may be expected for at least two reasons: (1) the mode filter itself may cause some mode conversion, and this mode conversion component will be influenced by the piston setting, and (2) the attenuation introduced by the mode filter is not sufficient to completely eliminate the wave components in the unused mode.

Even in the presence of the best mode filter now available, row 4 of Fig. 15 shows that the conversion-reconversion phenomenon does take place. The reconverted pulse amplitudes show up at 7,000 yards in row 4 as small distortions on the right hand side of the signal pulse, and these distortions grow in magnitude as we proceed to the right in the series of pictures representing more trips past the conversion points.

In the 5" line there are several modes for which the delay factor $z(1/v_x - 1/v_1)$ does not result in a separation between the reconverted pulse and the signal pulse until z is on the order of 5,000 feet. Thus, we might expect the conversion-reconversion phenomenon to broaden the signal pulse. This does indeed take place even for the favorable piston setting of the bare waveguide, as shown by the top line of pulses in Fig. 15. The pulses at the distance 3,500 yards are sharper than those pulses at the distance 27,700 yards. However, addition of the mode filter (which introduces negligible signal attenuation) does appreciably sharpen the pulse at the 27,700 yard distance (row 3).

Thus, on the basis of pulse transmission observations on the 5" line and a simple theoretical analysis, we conclude that the conversion-reconversion phenomenon will be important in a waveguide system, and that it is important to have as much dissipation as possible present in the unused modes of propagation.

ANALYSIS FOR CONTINUOUS MODE CONVERSION

The traveling pulse type of theoretical analysis utilized in the preceding section can be extended to describe a more realistic spatial distribution of conversion points and to include a series of unused modes instead of only one. An extension of this type is required in order to calculate directly the behavior which might be expected in a waveguide composed of randomly disposed irregularities.

A much simpler mathematical treatment, originally suggested to the writer by J. R. Pierce, is to assume uniform mode conversion along the direction of propagation and to represent this condition by a differential equation. An analysis of this type is attached as an appendix. The work includes the assumption of quadrature addition of conversion and reconversion components, and the total magnitude of such components given by the analysis may be thought of as the rms average of the conversion magnitudes in waveguide lines containing randomly located imperfections. Any single line might show somewhat more or less conversion effects, a factor of ± 10 db probably being adequate to cover most lines containing randomly located imperfections. If practical lines show appreciable correlation between the spacings of the conversion points, the reconverted-wave magnitude would become greater. The analysis has the advantage of being simple and understandable and should give overall trends accurately.

This analysis shows that the waveguide performance with regard to conversion-reconversion effects is completely specified with knowledge

of (1) the ratio of the conversion coefficient (a_{1x}) to the true dissipation coefficient* in the signal mode (a_{1h}), (2) the ratio of the heat-loss coefficient of the unused-mode to that for the signal-mode (a_{xh}/a_{1h}), and (3) the length of the transmission line specified in terms of decibels of heat loss to the signal wave. For a given ratio of conversion-loss to heat-loss, the same ratio of signal-power to reconverted-wave power will be present for a long low-loss waveguide as for a short high-loss waveguide. This makes it important to determine the ratio of conversion-loss to heat-loss for waveguides of several nominal attenuation coefficients and to predict these effects theoretically insofar as it is possible.

Another result of this analysis is plotted in Fig. 20, which shows the ratio of the signal power to the power in the unused mode at the end of the line, with transmission-line heat loss as the abscissa. These curves have been plotted for a fixed magnitude of conversion loss coefficient (a_{1x}) equal to 50 per cent of the true heat loss coefficient (a_{1h}) and for ratios (a_{xh}/a_{1h}), heat loss in the unused mode to heat loss in the signal mode, between 2 and 100. These values are typical of solid round waveguide without mode filters. It is interesting to note in Fig. 20 that the magnitude of the unused mode power relative to the signal mode power reaches very nearly a constant value in a transmission line length of only $1\frac{1}{2}$ to 1 db, except for extremely low ratios a_{xh}/a_{1h} . Physically what is happening is that the unused mode power becomes dissipated through heat loss about as rapidly as it is created by mode conversion, after an initial short transmission line length.

Fig. 21 shows the ratio of signal power to reconverted wave power as a function of transmission line length for the same conditions described in connection with Fig. 20. A heat loss ratio on the order of 2 to 10 is typical of important modes in solid round waveguide without the addition of mode filters,† and Fig. 21 shows that the ratio of signal-to-reconverted wave power for such a medium becomes poorer than 20 db for transmission line lengths longer than 1.5 to 2 db. Although there is some uncertainty as to the precise interpretation which may be placed on the signal power to reconverted wave power calculated in this manner, since the time relations in connection with a definite modulation method are not included, it seems evident that a solid copper tube without mode

* There is a very significant difference between the effects of signal power loss to other modes through conversion and signal power loss due to dissipation in the waveguide walls. However, it does not matter here whether the latter be due to surface roughness, chemical impurity or just the theoretical minimum heat loss for ideal copper. Therefore, all of the heat loss effects are combined into the single coefficient, a_{1h} .

† See the appendix for further discussion.

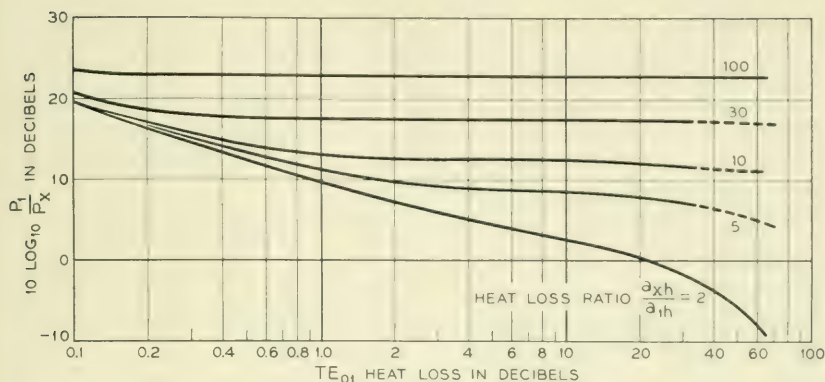


Fig. 20 — Ratio of TE_{01} signal power (P_1) to X-mode power (P_x) versus line length for conversion coefficients (a_{1x} and a_{x1}) equal to one-half the heat loss coefficient (a_{1h}).

filters is very unlikely to be satisfactory for long distances as a communication medium. However, the addition of mode filters will raise the heat loss to the undesired waves, and the latter improves the signal to reconverted-power ratio directly as the ratio of heat loss in the unused wave to heat loss in the signal wave. Thus, as shown on Figure 21, a heat loss ratio on the order of 500 would produce a ratio of signal power to reconverted wave power on the order of 20 db at a transmission line length of 60 db. It may be shown that the magnitude of the signal to reconverted wave power varies as the square of the conversion to heat loss coefficient ratio a_{1x}/a_{1h} .

The sharp break downward on the right hand end of the curves for $a_{xh}/a_{1h} = 2$ in Figs. 20 and 21 represents the condition wherein the power in the reconverted wave becomes comparable to the power remaining in the signal wave.

We next consider the improvement in signal fidelity which results from the introduction of ideal mode filters. We shall assume the ideal mode filters have a matched impedance for all modes, very high transmission loss to the unused modes, and no transmission loss for the signal mode. The improvement in the ratio of signal power to reconverted wave power due to the addition of such filters is shown in Fig. 22. This plot has been calculated for a total line length of 20 db heat loss, but the conclusions are valid for any line length wherein the signal wave power remains appreciably larger than the reconverted wave power. We observed that mode filters improve the signal-to-reconverted-wave powers very slowly when a_{1x} far apart, typical improvements ranging be-

tween 2 and 8 db for 1 db separation between mode filters. The larger improvement is obtained when the heat loss in the undesired mode is more nearly comparable to the heat loss in the signal mode before the addition of the mode filter. For very small spacing between the mode filters, addition of the ideal mode filters improves the signal-to-reconverted wave power by very large factors.

A somewhat different form of effect due to the addition of mode filters is observed if the line before the introduction of the mode filter has a reconverted wave power larger than the signal power. Under this condition, relatively large mode-filter spacings bring about a large improvement in signal-to-reconverted wave power, but this is due to the very poor condition present before filtering. It is doubtful whether the transmission line would become very useful without rather strong mode filtering of the type represented in Figure 22.

We may conclude that the mode filters should be placed very close together, preferably at spacings of less than .1 db heat loss to the signal wave. We may also conclude that the transmission of signals over distances corresponding to the order of 60 db heat loss will require either

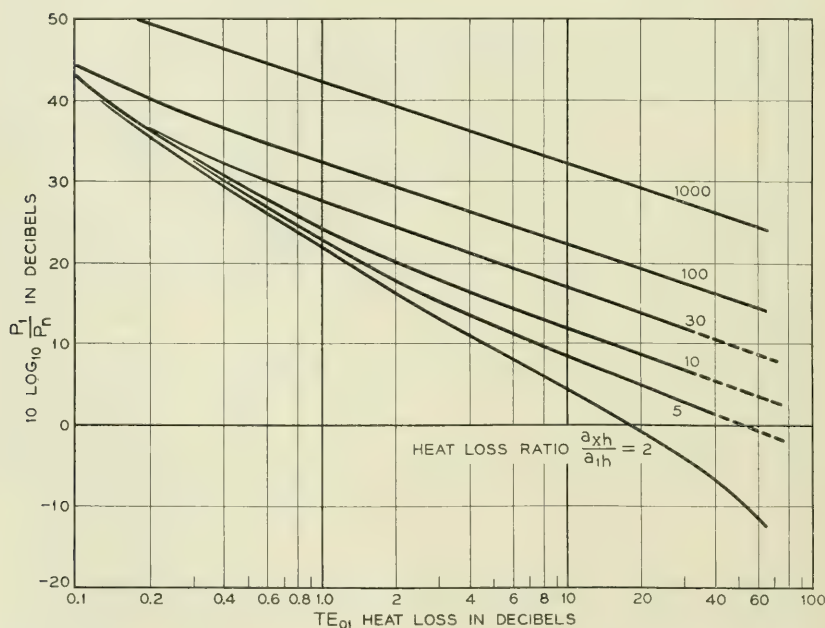


Fig. 21 — Ratio of TE_{01} signal power (P_1) to reconverted TE_{01} power (P_n) versus line length for $a_{1z} = 0.5 a_{1h}$.

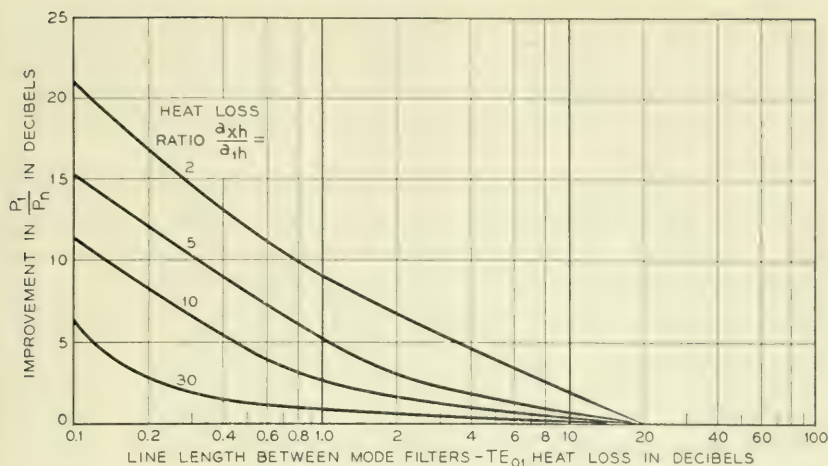


Fig. 22 — Improvement in P_1/P_n due to addition of mode filters. Total line length equals 20 db TE_{01} heat loss. (Plotted for $a_{1x} = 0.5 a_{1h}$).

the spacing of mode filters at less than 0.1 db to the signal wave or use of a continuous form of transmission line having a ratio of heat loss in the unused mode to the heat loss in the used mode on the order of 500.

DIRECT EVALUATION OF MODE CONVERSION MAGNITUDES

The important influence that mode conversion effects are expected to exert on signal fidelity lead us to make direct evaluations of the conversion coefficients. The direct evaluation consisted of transmitting (actually or in imagination) a pure circular-electric wave into one end of a waveguide section and, by measurement or by calculation on the basis of known geometry, determining the relative magnitude of the power converted to the unused modes.

The simplest experimental technique for analyzing mode impurities consists of a short radial probe at the guide wall. The radial probe responds to energy in any mode of propagation except the circular electric family, and serves as a versatile instrument for measuring the *order of magnitude* of mode conversion effects. The limitations of the technique stem from (1) the fact that the probe responds to the vector sum of the amplitude of the radial electric field components of about 35 modes (in the 5" line case), and this sum varies with circumferential and longitudinal position of the probe even though the power present in the modes is constant; and (2) the fact that the sensitivity of the probe response to a given magnitude of power in the guide is variable from mode to mode,

being (in an extreme case) 25 db greater for TE_{91} than for TE_{13} in the 5" pipe. For the majority of modes, however, the latter variation is ± 3 db, and the maximum probe response as a function of circumferential and (to a limited extent) longitudinal position can be determined without excessive labor.

The probe technique of mode conversion evaluation was first applied by M. Aronoff to the individual sections of the 5" diameter experimental line. He found that the average indication of conversion for the (approximately) 20 ft. lengths of pipe was 29.5 db below the signal wave power; since the power loss due to dissipation in the walls for a 20 ft. section is about 27 db below the signal wave power, the individual pipe measurements gave an order-of-magnitude estimate of 0.55 for the ratio of conversion loss to heat loss ($a_{1x} a_{1h}$). Four mechanically distorted sections of line, previously considered satisfactory, were identified and discarded on the basis of this approach.

A. C. Beck and M. Aronoff next applied the probe technique to the 5" diameter line assembled into lengths of 145 feet, 270 feet, and 500 feet. The indications of converted power were -17 db, -10.5 db, and -13 db respectively (at wavelengths near 3.2 cm) which is compatible with the hypothesis of random addition of a number of conversion components. Since the heat-loss power for the 500-foot line is about -13 db compared to the incident signal power, the 500-foot line radial probe measurement gave an order of magnitude estimate of 1.0 for $a_{1x} a_{1h}$, in fair agreement with the value of 0.55 from single pipe measurements. The probe indication of conversion as a function of frequency for the 500-foot line is plotted in Fig. 23, which shows that quite a number of

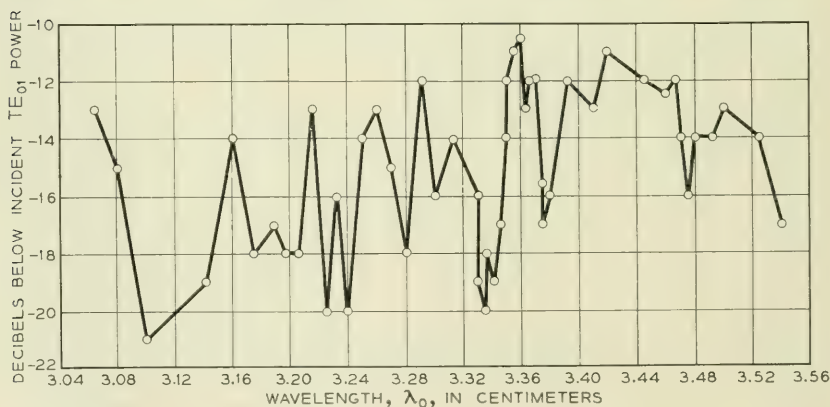


Fig. 23 — Probe recording of converted power in the 500-foot line.

conversion components were present; similar data were observed on the shorter lengths of line.

Azimuthal distributions of probe response showed conclusively that very little conversion was present to the 13 $TE_{n,m}$ and $TM_{n,m}$ modes having an index " n " of four or more.

A much more precise though more elaborate method of evaluating the conversion coefficients involves use of coupled wave transducers.² Such devices respond to only one mode and have known sensitivities, and therefore permit truly quantitative measurements. (At the present time this approach requires a separate transducer for each mode to be evaluated, a somewhat cumbersome procedure, but in principal the mode transducers can be made "tunable" for a series of modes.) A. C. Beck and M. Aronoff applied this more accurate method of conversion coefficient determination to several modes of the 500-foot line, and Table I shows the values averaged over the frequency band.

TABLE I — AVERAGE RATIO OF CONVERSION TO TE_{01} HEAT LOSS WITH TE_{01} EXCITATION

Mode	a_{1k}/a_{10}
TE_{11}	0.21
TM_{11}	0.05
TE_{21}	0.14
TE_{21}	0.05
TM_{01}	<0.001
Total	0.47

The estimated absolute accuracy is between 10 per cent and 20 per cent for these ratios. The variation of the conversion coefficients as a function of frequency is shown in Fig. 24 for two of the important modes and for the total of the modes given in Table I. The total of the modes measured is consistent with the probe indications, though the accuracy of the latter is low enough that this should not be interpreted as proving there are no other important conversion contributors.

The above direct evaluation of mode conversion in the 500-foot line yielded magnitudes that are sufficient to explain the conversion-reconversion process as already outlined. We wished to extend our experience with this particular line to higher-frequency lines which might be built and to absolute tolerances that might be placed on the construction of a new line. Toward this end, theoretical relations were derived by S. P. Morgan, Jr., for the mode conversions to be expected due to waveguide ellipticity, and due to the tilt and offset which may be expected to occur at the junction of two sections. Experimental work was done by M.

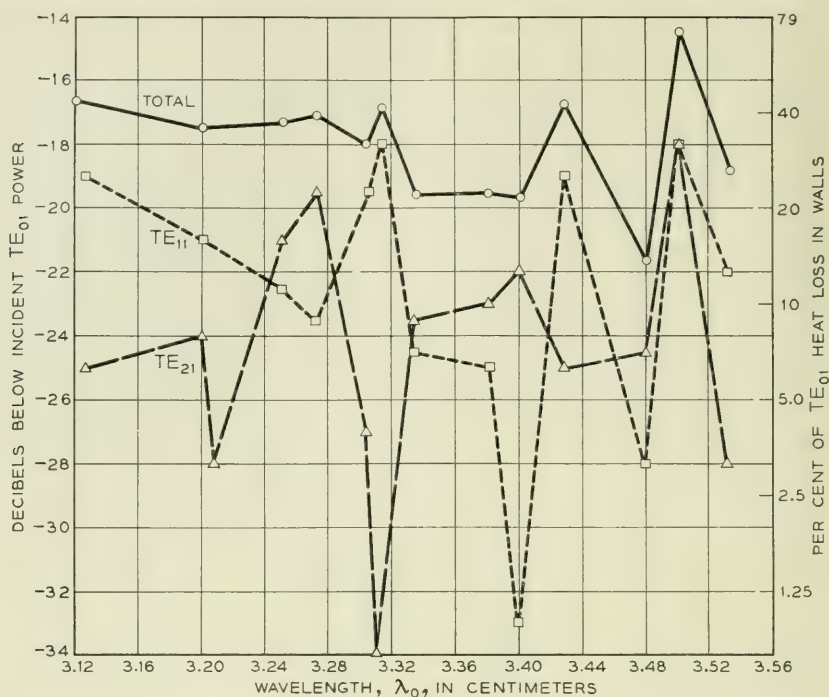


Fig. 24 — Observed conversion from TE_{01} to TE_{11} , TE_{21} , and the sum of TE_{11} , TE_{21} , TM_{11} , TM_{01} , and TE_{31} for the 500-foot line.

Aronoff³ at 9000 mc on accurately created imperfections of the above types, and he found excellent agreement. We proceed to use the theory to compare the present line to a hypothetical 50,000 mc line. The results are given in Table II.

These computations represent a single oval section or a single tilted or offset joint. The converted power varies as the square of the tilt angle, as the square of the offset distance, and as the square of the difference between the major and minor diameters. The total ratio of converted power to heat loss power depends on the number of conversions per unit length. For the accuracy of constructing a 2" diameter line assumed in Table II, the amount of mode conversion to be expected is not appreciably different from what appears attributable to known mechanical imperfections in the 500-foot line already discussed.

Another waveguide property of interest is the way the mode conversion magnitudes vary across the frequency band in a fixed pipe, and Table III shows this for the 2" diameter line.

TABLE II — MODE CONVERSION COMPARISON OF 9000-MC 1.732" DIAMETER LINE WITH A HYPOTHETICAL 50,000-MC 2" DIAMETER LINE

Imperfection		Pipe Diameter	Frequency	Magnitude of Converted Power	
Type	Magnitude			Percentage	Mode
Ovality	16* mils	inches 4.732	Mc 9000	0.53% [†]	TE ₃₀₁
Ovality	4*	2.0	50,000	0.18% [†]	TE ₂₀₁
Tilt	1°	4.732	9000	0.34%	TE ₁₁₂
Tilt	1°	2.0	50,000	2.0%	TE ₁₁₂
Offset	10 [†] mils	4.732	9000	0.008%	TE ₁₁₂
Offset	2.5 [†]	2.0	50,000	0.003%	TE ₁₁₂

* Difference between major and minor diameters.

[†] Separation of guide axes.

‡ These are upper-limit values, based on the length of the oval section of pipe which would produce maximum mode conversion, and based on a cross sectional shape (trifoil) which would produce the maximum of mode conversion.

It is interesting that two of the three conversion effects are essentially independent of frequency. Tilt at a waveguide junction introduces a phase-front error and would be expected to cause greater conversion effects at increasing frequencies. We shall see that bends produce a similar mode conversion, also due to a phase front error, that increases with increasing frequency.

THE BEND PROBLEM

The problem of transmitting the circular electric wave around bends was recognized as being important at an early date, and contributions to its solution were made by M. Jouguet,^{4,5} W. J. Albersheim,⁶ S. G. Rice,⁷ and the writer.⁸ The essence of the problem is as follows: A bend in a

TABLE III — FREQUENCY VARIATION OF MODE CONVERSIONS IN 2" DIAMETER GUIDE

Imperfection		Mode	Magnitude of Converted Power		
Type	Magnitude		$f = 24,000$ mc	$f = 50,000$ mc	$f = 75,000$ mc
Ovality	4 mils	TE ₃₁	0.22%*	0.18%*	0.20%*
Tilt	1°	TE ₁₁₂	0.41%	2.0%	4.8%
Offset	2.5 mils	TE ₁₁₂	0.003%	0.003%	0.003%

* These are upper-limit values, based on the length of the oval section of pipe which would produce maximum mode conversion, and based on a cross-sectional shape (trifoil) which would produce the maximum of mode conversion.

round guide causes coupling between the circular electric wave (TE_{01}) and the TM_{11} wave, and in round solid pipe the TE_{01} and TM_{11} waves are degenerate, i.e., they have identical phase constants. This degeneracy has the effect of bringing in phase all components transferred to TM_{11} no matter how gradually the bend might take place. Theory neglecting dissipation^{4,7} shows that in a round pipe bent with any radius of curvature, the power will flow from the TE_{01} mode to the TM_{11} mode and back as a function of the total bend angle θ according to the relations

$$TE_{01} \text{ amplitude} = \cos\left(\frac{\pi \theta}{2 \theta_c}\right) \quad (9)$$

$$TM_{11}'' \text{ amplitude} = \sin\left(\frac{\pi \theta}{2 \theta_c}\right) \quad (10)$$

where

$$\theta_c = \frac{\pi}{2.32} \frac{\lambda_0}{a}$$

λ_0 = free-space wavelength

a = waveguide radius

A solid round pipe is unsatisfactory for transmission of the circular electric wave around bends in the broad bands we seek to use.

One method of making the guide satisfactory in bends is to break the degeneracy between the TE_{01} and TM_{11} waves. Use of an elliptical pipe has been shown theoretically to be one way of doing this. For a 2" diameter guide at 50,000 mc an eccentricity of about 0.3 permits a bending radius of 700 feet with theoretical bend losses in the range 0 to 0.17 db for any total bend angle; the heat loss coefficient for such an elliptic guide is about 35 per cent higher than for a perfectly round guide.⁸

Another method of avoiding bend losses is to introduce dissipation to the TM_{11} wave without adding loss to the TE_{01} wave. It has been shown² that a large difference between the attenuation coefficients of two coupled waves reduces the power transferred from the low-loss wave to the high-loss wave. Applied to the bend problem, this means that a structure with increased TM_{11} loss may be bent with less signal (TE_{01}) loss even though the phase constants might be degenerate. The reader is referred to the earlier paper⁸ for a more complete discussion. There exist several alternate forms of circular electric waveguide (to be discussed) which have an attenuation coefficient for TM_{11} more than 5,000 times the attenuation coefficient for TE_{01} . The calculated extra loss in the bend region for such structures and for solid round pipe has been plotted

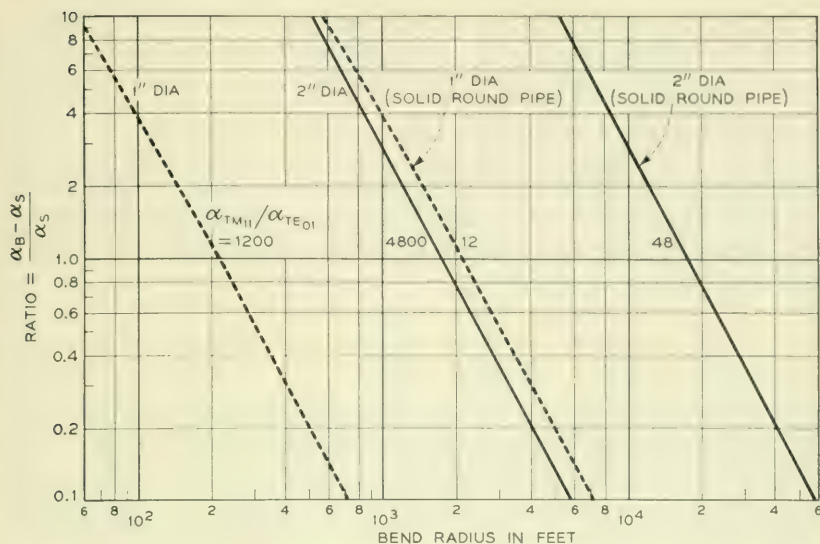


Fig. 25 — Computed change in the 50,000 mc TE_{01} attenuation coefficient due to bends in 1" and 2" diameter solid pipes and in modified guides having TM_{11} attenuation coefficients larger (than in solid pipe) by a factor of 100. α_B and α_s are the bend-region and straight-line attenuation coefficients, respectively.

as a function of bending radius in Fig. 25, assuming the TM_{11} - TE_{01} coupling due to the bend is the same in the altered guide as in solid round pipe. Whereas a bending radius of 17,500 feet causes a 100 per cent increase in TE_{01} heat loss for 50,000 mc waves in 2" diameter solid pipe, the modified structure with a TM_{11} attenuation coefficient that is larger by a factor of 100 should tolerate a bending radius of 1,750 feet for the same heat loss increase. (The 50,000 mc attenuation coefficients for ideal 1" and 2" copper pipes are 14.8 and 1.79 db/mile respectively.)

For estimating purposes, the ratio of the extra loss per unit line length in the bend region to the straight line loss may be calculated for solid round pipes from the relation

$$\frac{\alpha_B - \alpha_s}{\alpha_s} = \frac{2.5 \times 10^{10} a^6}{\lambda_0^3 R^2} \quad (11)$$

and the absolute increase in attenuation due to a bend is*

$$(\alpha_B - \alpha_s) = \frac{9.7 \times 10^5 a^3}{\lambda_0^{3/2} R^2} \text{ (db/meter for copper guide)} \quad (12)$$

* For small bend radii and bend angles less than θ_c , this relation gives a greater loss than the correct value. See Figs. 22 and 23 of Reference 8 and also see Reference 2.

where α_s = straight line attenuation coefficient

α_B = bend region attenuation coefficient

a = guide radius

R = bending radius

λ_0 = free space wavelength

The approximations used in deriving (11) and (12) are good when the operating frequency is at least 50 per cent greater than cut off for TE_{01} . For guides modified to have higher TM_{11} attenuation both (11) and (12) may be divided by the factor

$$\frac{\alpha_{TM_{11}}^m}{\alpha_{TE_{01}}^m} \cdot \frac{\alpha_{TE_{01}}^0}{\alpha_{TM_{11}}^0} \quad (13)$$

where α^m and α^0 denote attenuation coefficients for the modified guide and solid round guide respectively. On the assumption that the mode coupling is the same in the modified guide as in the solid round guide, use of (13) with (11) or (12) provides an estimate of bend losses in modified circular-electric waveguides.

For a fixed ratio of bend-region attenuation to straightline attenuation, the allowable bending radius varies inversely as the square root of the ratio of TM_{11} heat loss to TE_{01} heat loss, varies inversely as $\lambda_0^{3/2}$, and varies directly as the third power of guide radius.

For fixed bending radius, the *absolute* bend loss varies inversely as $\lambda_0^{3/2}$; since the straight line TE_{01} loss varies directly as $\lambda_0^{3/2}$, bend losses tend to equalize the overall heat loss versus frequency characteristic of the waveguide.

IMPROVED FORMS OF CIRCULAR ELECTRIC WAVEGUIDE

In the preceding discussion it has been indicated that added dissipation for the unused modes of propagation has the effect of decreasing signal losses and of reducing the interference effects associated with mode conversion. Dissipation can be introduced to the unused modes of propagation through the addition of mode filters at intervals along the line, but it appears very desirable to introduce the dissipation to the unused modes on a continuous basis. Several ways of making the line lossy to the non-circular electric modes have been found, and one is illustrated in Figure 26. The copper rings lie in planes transverse to the direction of propagation and provide the conductivity required as a boundary for the circular electric wave family. Successive rings are insulated from each other, however, and the guide provides very poor conductivity in the longitudinal direction. All modes other than the

circular electric wave family have wall currents in the longitudinal direction and experience considerably increased loss in the spaced-ring structure compared to a solid-walled waveguide. A. G. Fox first observed that the spaced ring structure could be used to transmit circular electric waves around bends, and since that time additional work has been carried out by A. P. King and M. Aronoff. The observed loss for the spaced-ring structure under proper conditions was observed to be about 60 per cent more than the theoretical loss for an ideal copper tube, whereas the observed loss for the unused modes of propagation was on the order of 1,000 to 5,000 times the circular electric wave value.

The spaced-ring structure therefore has the electrical properties we seek. The higher-order circular waves exist with losses comparable to their values in a solid copper pipe, but fortunately the magnitudes of conversion between the waves of the circular-electric family have been found to be small. The spaced-ring structure does present some difficult problems with regard to fabrication.

An analogous structure composed of a continuous helical conductor supported within a lossy housing has electrical properties which approximate those of the spaced-ring structure, and the helix should be considerably easier to manufacture in long lengths. The helix might be expected to support a wave-type approximating the circular electric wave both from the standpoint of field distribution and loss when one observes that a helix of very small pitch presents almost circumferential conductivity as required by the circular-electric wave, and the very small longitudinal component necessary due to the finite wire size tends toward zero as the helix pitch tends toward zero. James A. Young of these laboratories has constructed helices in the 2 db/mile waveguide size (4.73" diameter at 9,000 mc) and found a heat-loss coefficient on the order of 1.75 times the theoretical value for ideal copper pipe. These large ex-

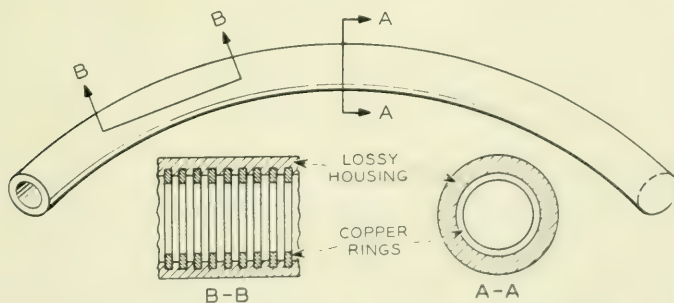


Fig. 26 — Spaced-ring circular electric waveguide.

perimental helices were known to be imperfect, and in a smaller diameter of helix, losses within 15 per cent of the theoretical values for perfect copper pipe were achieved. These results lead us to regard the helical line as a very promising medium for circular electric waves.

SURFACE ROUGHNESS

Since frequencies on the order of 50,000 mc are desirable for low-loss waveguide use, it was recognized that roughness of the surface at the waveguide walls might appreciably increase the heat loss in a practical waveguide. The first approach to this problem was made by W. A. Tyrrell using sections of 5 diameter copper pipe from the experimental line. Tyrrell measured the heat-loss coefficients of the pipe when used as a resonant cavity at 9,000 mc in lengths on the order of 4 to 8 feet. Carefully selected resonant conditions were employed to avoid bringing the unused modes of propagation into resonance at the same time that the circular electric wave was resonated. Whereas Tyrrell observed that the heat loss coefficient in the pipe as originally drawn was about 21 per cent higher than the value computed using the measured dc conductivity, he found that rotary* grinding and polishing the inner surface of the guide reduced the excess loss to about 12 per cent. Tyrrell also observed that commercially drawn brass and 2S aluminum tubing of approximately the same dimensions showed measured losses 11 per cent and 20 per cent respectively greater than the value predicted from the measured dc conductivity. Therefore, the indication from Tyrrell's work was that surface roughness did indeed account for increased losses even at 9,000 mc, and that the excess losses could be reduced either by polishing the surface or through the use of lower conductivities (which have the effect of increasing the skin depth).

A parallel approach to the measurement of surface-roughness effects was made by A. C. Beck and R. W. Dawson, also at 9,000 mc.⁹ Beck and Dawson used small wire samples as the center conductor of a coaxial cavity† and found that commercially drawn copper, aluminum and silver wires showed loss values 10 per cent and 15 per cent higher than those expected from the measured dc conductivity. By mechanically polishing

* Because the wall currents for the circular electric wave are circumferential, the longitudinal surface scratches produced by drawing are in the worst possible orientation. Polishing was carried out in a rotary manner so that the current would not cross the scratches so induced.

† In a coaxial, the currents are longitudinal, as are the scratches from drawing, so the measurements in the coaxial would be expected to show somewhat less excess loss due to surface roughness than the measurements made in circular-electric waveguide cavities.

these same wires the losses were reduced to 5 to 8 per cent above the de values and, by electropolishing, copper wires were brought within 2 per cent of the theoretical value.

An investigation of the effects of aging on the 9,000 mc conductivity of copper surfaces was conducted jointly by Dawson, Tyrrell and Beck, using wires which were measured in the coaxial cavity.⁹ They found (1) that the conductivity of untreated surfaces remained essentially stationary when stored indoors; (2) that the conductivity of bare electropolished surfaces deteriorated (over a period of months) to a value comparable to that for commercially drawn wire; and (3) that a tight bonding coating very greatly slows down the aging process.

Recent measurements made in the vicinity of 50,000 mc on a sample of commercially-drawn fine-silver waveguide have indicated that surface roughness effects increased the loss values by approximately 20 per cent.

The overall conclusion would appear to be that it is now feasible commercially to produce waveguide surfaces which have excess losses due to surface roughness no greater than 20 per cent even at frequencies near 50,000 mc, and that by refining the manufacturing method it may be possible to approach the ideal value more closely. In order to avoid aging effects, a tight-bonding coating and a protective atmosphere may be desirable.

CIRCULAR ELECTRIC WAVE AND MILLIMETER WAVE TECHNIQUES

A whole new line of components and techniques are required to carry out in the millimeter wave region the same functions that have always been required in a system application, i.e., power generation, amplification, frequency-band separation, hybrid division, amplitude and phase equalization, and so forth. With a view to ultimate use in a waveguide transmission system, basic research has been done on all of these elements over a period of years at these Laboratories.

The importance of primary oscillators and amplifiers to millimeter-wave systems is self-evident, and work has begun going forward in this difficult field under the direction of J. R. Pierce. The results have been published during the last several years. Pierce made the first 6 mm reflex klystron oscillator.¹⁵ The first 6-mm amplifier, made by John Little,¹⁰ was a travelling wave tube using a conventional helix and it provided 3 db gain. Later, S. Millman¹¹ introduced the space-harmonic type of travelling-wave amplifier and built several which had over 20 db gain near 6 mm. R. Kompfner¹² devised the backward-wave type of travelling-wave oscillator and built a tube which oscillated from about 5 mm to 8 mm. A. Karp¹³ devised a simple structure for backward-wave

oscillators which has already resulted in 5 to 6 mm oscillators and which appears suitable for use at shorter wavelengths.* Recently, E. D. Reed produced a 5 to 6 mm reflex klystron.*

R. S. Ohl has continued his pioneering work on point-contact crystal converters and has made units for use at 6 mm having conversion losses of less than 8 db and output noise ratios on the order of three times basic thermal noise. Ohl also made point contact silicon units for use as harmonic generators to permit the conversion of 24,000 mc power to 48,000 mc power. His harmonic generators have proven invaluable as a source of millimeter wave power — essentially all of the radio research work done to date has been carried out using them.

In order to evaluate crystals and millimeter wave oscillators, it is essential to have an absolute power reference in the millimeter region, and work* has been done by W. M. Sharpless to establish such a reference.

Up to the present time all of the amplifiers, oscillators and other circuit elements have employed dominant-mode rectangular waveguides in order to simplify the circuit design. Therefore, it is of importance to know how to transform a signal from a dominant-mode rectangular guide to the circular electric wave in round pipe. The first circular-electric-wave transducer made in these Laboratories was designed by A. P. King and had the form sketched in Fig. 27. This transducer is of the general type in which the metallic boundary of the waveguide is shaped to force the field lines in the cross section of the guide into the pattern characteristic of the desired output wave. In Fig. 27 the dominant-mode rectangular guide at the left end is gradually tapered to the sector of a circle; the size of this sector is small enough so that only one wave type can exist at this point, and the electric field lines are arcs of a circle. Next, the angle of the sector is gradually increased along the axis of propagation until at one point a cross section of the guide has the shape of a half circle. The size of the sectoral angle is continually increased, however, until finally the metallic sector of the circle disappears as a radial vane. When the taper is done gradually (an overall length of approximately 10 to 15 wavelengths) the electric field lines remain arcs of a circle as

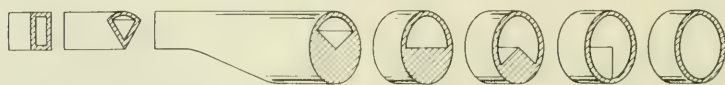


Fig. 27 — Circular-electric wave transducer (due to A. P. King).

* A portion of this work was carried out under Office of Naval Research Contract Nonr 687(00).

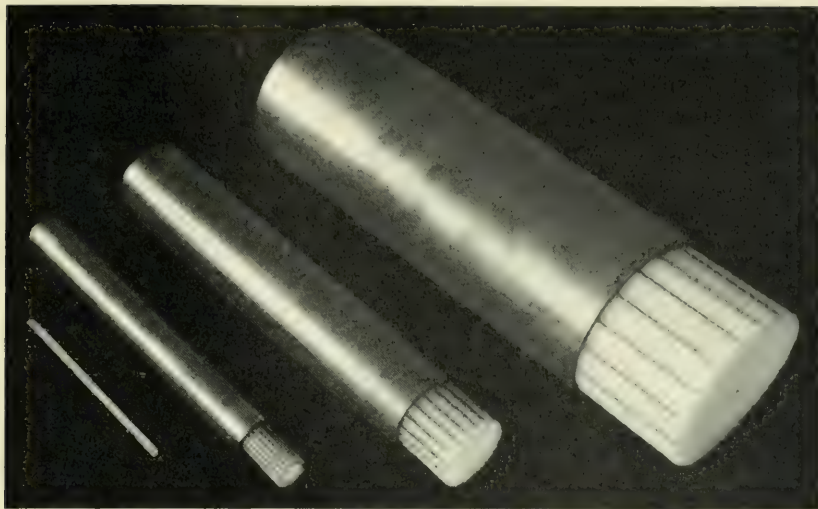


Fig. 28 — Mode filters which pass only circular electric waves.

they were in the sector at the left-hand end of Fig. 27, and the circular-electric wave emerges in the round pipe. This type of transducer has been shown to have transfer losses from dominant-mode rectangular guide to circular-electric wave in round pipe of approximately 0.3 db at 24,000 mc. Similar models have been made by A. G. Fox for use at 48,000 mc.

Another important component is the mode filter previously referred to and which attenuates all wave types other than the circular electric wave family. One type of mode filter to perform this function is the spaced ring structure of Fig. 26 and another type, due to A. P. King, consists of resistive sheets along radial planes as shown in the photograph of Fig. 28. The circular electric wave family has no electric field in a radial direction or in a longitudinal direction. All other wave types, however, have radial-electric or longitudinal-electric field components and experience attenuation due to the presence of the resistive sheets.

The coupled-wave type of transducer sketched in Fig. 29² is useful in connecting from dominant rectangular guides to the circular electric or other modes in round guide. This type of transducer makes use of the fact that the various modes in the multimode guide have unequal phase constants. The transfer of power from rectangular guide to the round guide takes place only to the particular round-guide mode whose phase constant is equal to that of the wave in the rectangular guide. This type

of transducer has the same geometric appearance (aside from exact dimensions) for any mode in the round guide and is attractive in that it presents a matched impedance to all the modes of propagation. This property may be used to combine a series of signals onto different modes in a single transmission line. The coupled wave transducer may also be employed to multiplex a series of frequency bands into one pipe. We may wish to employ the long-distance waveguide over the frequency band from perhaps 35,000 mc to 75,000 mc and will require a series of transducers to go from dominant mode guides in various portions of this band to the circular-electric wave in the round guide. Frequency-selective coupled-wave transducers may be employed in the manner sketched in Fig. 30 to multiplex these frequency bands into the pipe for the long distance transmission.

A. G. Fox¹⁴ has shown that dielectric waveguides are attractive as a flexible connecting link for terminal equipment in the millimeter wave region and may also be employed in circuits such as hybrids.

On all of these items of millimeter wave technique and multimode waveguide technique, individual publications will appear as soon as the work has reached the point where this becomes possible.

MODULATION METHODS

The modulation method to be used for the transmission of intelligence on a waveguide system will probably be dominated by the conversion-reconversion phenomenon already discussed. In order to evaluate the

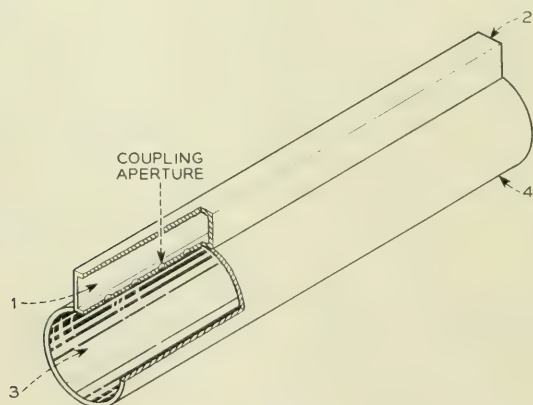


Fig. 29 — Coupled-wave transducer for generating circular-electric or other waveguide modes.

interfering effects of the conversion process, it is important to take into account the time relations between the signal components and the reconverted wave components. In general, it seems clear that reconverted energy returning to the signal mode within a time interval appreciably less than the reciprocal of the highest modulation frequencies will be less interfering than reconverted wave components separated from the signal components by time intervals on the same order as the reciprocal of the modulation frequencies. In practice, a distribution of reconversions will be obtained from a line containing randomly disposed imperfections. As discussed in connection with equation (1), an approximate upper limit can be placed on the time separation between the signal energy and the reconverted wave energy. Table IV records the "upper-limit" time delay for the 5" diameter experimental 9000 megacycle waveguide

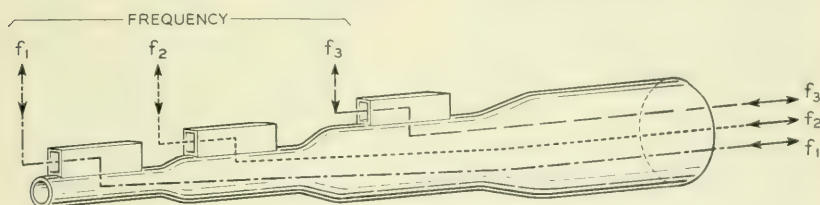


Fig. 30 — Proposed method of multiplexing several frequency bands in the circular electric waveguide.

and for a 2" diameter pipe used at a frequency of 48,000 mc. The negative time delays represent reconverted-wave components which precede the signal at the receiving end of the system. These calculations suggest that reconverted wave energy might precede the signal by almost $\frac{1}{2}$ microsecond or lag the signal by approximately 1 microsecond in the 5" diameter 9,000-mc line. Photographs of the observed pulse transmission shown in Fig. 15 indicate negligible reconverted-wave energy at more than one-half these time intervals, and therefore it appears that the method of estimating (described in connection with equation 1) is pessimistic. However, the "upper limit" time delay does provide a convenient way of comparing the experimental line with a proposed 2" diameter line used at 48,000 mc. Observe in Table IV that the 2" diameter line will show appreciably less delay for the reconverted wave components than is to be expected from the 5" diameter line. This is due to higher attenuation in the unused modes and to smaller differences between the group velocities of the various modes in the 2" diameter pipe. "Upper limit" time delays in the range 1 to 100 millimicroseconds

TABLE IV — "UPPER LIMIT" TIME DELAY FOR RECONVERTED ENERGY COMPONENTS IN SOLID PIPE WITHOUT MODE FILTERS

Mode	"Upper Limit" Time Delay	
	4.732" Dia. Pipe at 9000 Mc	2.0" Dia. Pipe at 48,000 Mc
TE ₁₁	-0.456×10^{-6} second	-0.0137×10^{-6} second
TE ₂₁	-0.097	-0.0035
TE ₃₁	+0.0357	+0.0014
TE ₁₂	1.093	+0.111

are indicated for the 2" diameter line and may also be regarded as pessimistic estimates by virtue of our experience with the 5" diameter line.

It is to be emphasized that the time delays of Table IV apply to solid round guide without mode filters. The addition of mode filters in the solid pipe would increase the loss for the unused modes and appreciably reduce the "upper-limit" time delays for reconverted wave energy. For the improved forms of circular-electric waveguide the increased losses for the unused modes would result in similarly reduced time delays; typical factors of reduction might run between 100 and 1,000 for spaced-ring and helical waveguides compared to the data for solid pipe given in Table IV.

Taking into consideration both conversion-reconversion effects and some equalization of delay-distortion due to waveguide cut-off, it would appear likely that baseband widths on the order of 100 to 1000 mc could be employed in a 2" diameter line.

The character of the reconverted wave interference will be as discussed in connection with Figs. 15 and 18 if the modulating wave of the signal carrier contains frequencies near the reciprocal of the "upper-limit" time delay for the particular waveguide used.

The magnitude of the reconverted wave energy has been discussed in connection with Fig. 21, and it has been observed that ratios of signal power to reconverted-wave power on the order of 20 db may be expected on a well engineered waveguide line having 60 db of heat loss for the signal wave. Although the time relations were not specifically taken into account in this discussion, it seems likely that the order of magnitude of the reconverted-wave power will be unaltered by more precise analysis.

One might therefore conclude that the modulation method to be used in a waveguide system must be one which will tolerate large amounts of signal interference. Pulse code modulation is one arrangement of signaling which will tolerate such interference, and undoubtedly there are others as well.

Another conclusion the writer has reached is that some form of signal regeneration is likely to be required at each repeater, and that the modulation process should be selected in such a way as to permit this regeneration with limited complexity.

CONCLUSION

Single-mode waveguide is unattractive as a long-distance communication medium due to limited bandwidth and either unreasonable size or excessive loss.

The circular-electric mode in a 2" diameter round pipe has a theoretical attenuation coefficient of 2 db per mile for carrier frequencies near 50,000 mc. Delay distortion due to waveguide cut-off will require equalization if baseband widths on the order of 500 mc are to be provided on repeater spacings of 25 miles. Using frequency-division multiplex, such a waveguide might be exploited over the 40,000-mc band from 35,000 to 75,000 mc, for which the theoretical attenuation coefficients are 3 db/mile and 1 db/mile respectively.

Transmission experiments were conducted at 9,000 mc in a 500-foot copper pipe 4.73" I.D. for which the theoretical circular-electric wave loss was 1.9 db/mile. Under favorable conditions the observed losses were within 25 per cent of the theoretical value; under unfavorable conditions (which are not likely to occur in practice) the observed losses were as high as 75 per cent greater than the theoretical value. Surface roughness accounted for losses about 20 per cent above the theoretical value, and the remaining excess losses were due to conversion of energy to other modes of propagation. Direct observation of the power transferred between modes in the 500-foot line confirmed the latter conclusion.

Other observations in the 500-foot line showed that the mode conversion process produced a signal-distortion or signal-crosstalk effect through reconversion of energy from the unused modes of propagation back to the signal (circular electric) mode. This type of interference seriously limits the capabilities of bare copper pipe for use as a long-distance communication medium. However, it has been shown experimentally and theoretically that the effects of the conversion-reconversion process are greatly reduced through the addition of mode filters, which absorb energy present in the unused modes of propagation.

The combination of solid pipe plus mode filters remains attractive as a communication medium. Average losses for the unused modes of propagation should be on the order of 500 times the loss to the signal wave, in order that the conversion-reconversion effects be tolerable in

the presence of conversion losses of the magnitude observed in the 500-foot experimental line.

Improved forms of circular-electric waveguide have been demonstrated. These waveguides have the common property of providing very large attenuation coefficients for the unused modes of propagation while maintaining essentially the same circular-electric wave attenuation as in solid pipe. Thus, these guides are essentially continuous mode filters. Structurally, such a medium may consist of a series of metallic rings supported in a lossy housing (Fig. 26), or may consist of a helix supported in similar manner. The helical circular-electric waveguide appears more attractive from the standpoint of ease of fabrication.

Where bends must be sharp, they may be negotiated in a number of special ways. Gradual bends may be introduced in the improved forms of circular-electric waveguide with modest increase in heat loss. For the 2" diameter guide at 48,000 mc, it is estimated that a bending radius of about 2,000 feet would double the heat loss of a helical or spaced-ring waveguide. For a 1" diameter helical or spaced-ring guide (theoretical heat loss of 15 db mile in solid pipe) the corresponding bending radius is about 200 feet. The extra heat loss due to bending varies inversely as the square of the bending radius.

The type of modulation to be used in a waveguide system will probably be dominated by conversion-reconversion effects. An upper limit on the time delay between the signal component and the associated reconverted-wave components lies in the range 1 to 100 millimicroseconds for 2" pipe at 48,000 mc without mode filters; the addition of mode filters or use of the helical circular-electric guide should reduce these time delays by factors of 100 or more. Thus, it appears likely that basebands on the order of 500 mc or more should be usable.

It is concluded that a waveguide signalling method must be capable of withstanding large amounts of signal interference, and that some form of regeneration is likely to be required at each repeater.

ACKNOWLEDGMENT

The encouragement of R. Bown, H. T. Friis, and J. C. Schelleng is gratefully acknowledged. The contributions of numerous colleagues (as noted throughout the paper) form the building blocks without which the present understanding of waveguide transmission could not have been obtained.

APPENDIX

THEORETICAL ANALYSIS FOR CONTINUOUS MODE CONVERSION

Whereas the travelling-pulse type of theoretical analysis utilized in the body of this paper can be extended to a realistic spacial distribution of conversion points and to a series of modes instead of only one, J. R. Pierce suggested that the assumption of uniform mode conversion along the axis of propagation would lead to a solution in closed form and would probably show the general properties being sought. This suggestion was adopted and P_1 is designated as the signal power, P_x as the power in the unused mode, and P_n as the power which has transferred from mode x back to the signal mode, mode 1. We assume quadrature addition of conversion components, and write a series of differential equations expressing the *power flow* between the modes along the axis of propagation, including the heat loss effects:

$$\frac{dP_1}{dz} = -a_{1h}P_1 - a_{1x}P_1 \quad (9)$$

$$\frac{dP_x}{dz} = -a_{xh}P_x - a_{x1}P_x + a_{1x}P_1 + a_{1x}P_n \quad (10)$$

$$\frac{dP_n}{dz} = -a_{1h}P_n - a_{1x}P_n + a_{x1}P_x \quad (11)$$

in which the symbols have the following definitions:

a_{1h} = the heat loss coefficient — mode 1

a_{1x} = the mode conversion coefficient from mode 1 to mode x

a_{xh} = the heat loss coefficient — mode x

a_{x1} = the mode conversion coefficient from mode x to mode 1

z = distance along the axis of propagation

Note that the above heat loss coefficients are those associated with power rather than attenuation coefficients associated with amplitudes, ($2\alpha_1 = a_{1h}$, $2\alpha_x = a_{xh}$). The above equations also imply mode conversion in the forward direction only.

In a phenomenological way, these equations represent the decay of power in the signal mode and the build up of power in both the unused mode x and reconverted energy P_n in mode 1. The general plan is to solve these equations for P_x and P_n in terms of the input wave power. P_n is maintained separate mathematically from P_1 , even though both of them are in the same mode, so that we can clearly identify the energy which has been at one time in the unused mode x .

For mathematical solution, (9), (10) and (11) may be put in the following form:

$$\frac{dP_1}{dz} + \alpha P_1 = 0 \quad (12)$$

$$\frac{dP_x}{dz} + \beta P_x - a_{1x}P_1 - a_{1x}P_n = 0 \quad (13)$$

$$\frac{dP_n}{dz} + \alpha P_n - a_{x1}P_x = 0 \quad (14)$$

in which $\alpha = a_{1h} + a_{1x}$

$$\beta = a_{xh} + a_{x1}$$

The general solution for (12), (13) and (14) is given by the following

$$P_1 = k_1 \epsilon^{-\alpha z} \quad (15)$$

$$P_2 = k_2 \epsilon^{r_1 z} + k_3 \epsilon^{r_2 z} \quad (16)$$

$$P_n = -k_1 \epsilon^{-\alpha z} + \frac{k_2}{a_{1x}} (r_1 + \beta) \epsilon^{r_1 z} + \frac{k_3}{a_{1x}} (r_2 + \beta) \epsilon^{r_2 z} \quad (17)$$

where

$$r_{1,2} = \frac{-(\alpha + \beta) \pm \sqrt{(\alpha - \beta)^2 + 4a_{1x}a_{x1}}}{2} \quad (18)$$

The positive sign is to be associated with r_1 and the negative sign with r_2 . For the boundary conditions, at $z = 0$,

$$P_1 = P_0$$

$$P_x = 0,$$

$$P_n = 0,$$

that is to say, the input to the transmission medium being zero in both the x mode and reconverted energy mode, then the solution takes the following form:

$$P_1 = P_0 \epsilon^{-\alpha z} \quad (19)$$

$$P_x = \frac{a_{1x}P_0}{(r_1 - r_2)} \epsilon^{r_1 z} - \frac{a_{1x}P_0}{(r_1 - r_2)} \epsilon^{r_2 z} \quad (20)$$

$$P_n = -P_0 \epsilon^{-\alpha z} + \frac{(r_1 + \beta)P_0}{(r_1 - r_2)} \epsilon^{r_1 z} - \frac{(r_2 + \beta)}{(r_1 - r_2)} \epsilon^{r_2 z} \quad (21)$$

It is informative to note that $(r_1 - r_2)$ is always positive and is equal to

$$\sqrt{(\alpha - \beta)^2 + 4a_{1x}a_{x1}}$$

We are usually interested in the ratio of the power in the x-mode to the signal power P_1 and the ratio of the reconverted energy P_n to the signal power P_1 . These ratios are given by the following expressions

$$\frac{P_n}{P_1} = \frac{(r_1 + \beta)}{(r_1 - r_2)} \epsilon^{(r_1 + \alpha)z} - \frac{(r_2 + \beta)}{(r_1 - r_2)} \epsilon^{(r_2 + \alpha)z} - 1 \quad (22)*$$

$$\frac{P_x}{P_1} = \frac{a_{1x} \epsilon^{(r_1 + \alpha)z}}{(r_1 - r_2)} [1 - \epsilon^{-(r_1 - r_2)z}] \quad (23)*$$

Thus, we have explicit solutions for the uniform transmission medium containing mode conversions.

In order to make the most general study of these relations, we shall express the mode conversion coefficients in terms of the heat loss coefficient in the signal mode—i.e., as the ratio a_{1x}/a_{1h} . This is natural enough physically, for we are interested in the relative magnitudes of the heat loss and mode conversion effects. It is found that knowledge of the ratios a_{1x}/a_{1h} , a_{x1}/a_{1h} and a_{xh}/a_{1h} enables us to completely determine P_x/P_1 and P_n/P_1 in terms of the distance parameter $\epsilon^{-a_{1h}z}$. The latter is the heat loss in the signal mode, another familiar physical characteristic.

CHARACTERISTIC CONDITIONS IN BARE ROUND WAVEGUIDE

In order to use the theoretical relations derived in the preceding section, we need to know typical values of the parameters. In particular, we need to know the magnitude of typical conversion coefficients a_{1x} and values of the heat loss coefficients a_{1h} and a_{xh} for the modes of interest.

One set of heat loss coefficients which is of immediate interest may be made up from the calculated values for the 5-inch diameter round waveguide used in waveguide experiments at Holmdel. Table V shows the ratio of attenuation coefficients for several modes in this line. The circular electric wave TE_{01} has the lowest attenuation coefficient (absolute value

* When using these relations, it is helpful to note that $(r_1 + r_2) = -(\alpha + \beta)$ at all times. Hence $(r_1 + \alpha) = -(r_2 + \beta)$ and $(r_2 + \alpha) = -(r_1 + \beta)$.

TABLE V — RATIO OF ATTENUATION COEFFICIENTS FOR SEVERAL MODES IN 4.732" DIAMETER. WAVEGUIDE AT $\lambda_0 = 3.0$ CM

Modes	Attenuation Coefficient Ratio α_x/α_1 or a_{xh}/a_{1h}	Modes	Attenuation Coefficient Ratio α_x/α_1 or a_{xh}/a_{1h}
TE ₁₂ /TE ₀₁	2.45	TE ₃₁ /TE ₀₁	12.5
TE ₀₂ /TE ₀₁	3.82	TE ₄₁ /TE ₀₁	17.
TE ₁₁ /TE ₀₁	4.56	TE ₅₁ /TE ₀₁	21.
TE ₂₂ /TE ₀₁	4.63	TE ₆₁ /TE ₀₁	27.
TE ₁₃ /TE ₀₁	6.61	TE ₇₁ /TE ₀₁	34.
TE ₂₁ /TE ₀₁	8.51	TE ₈₁ /TE ₀₁	44.
TM ₁₁ /TE ₀₁	10.78	TE ₉₁ /TE ₀₁	61.
TE ₀₃ /TE ₀₁	11.3	TE _{10,1} /TE ₀₁	100.

equal to 1.6 db/mile), and the a_{xh}/a_{1h} ratios for the other modes range from 2.5 to 100 times the TE₀₁ value. Table V represents a selection of the various modes which can exist in the 5-inch diameter line, but the number tabulated is *not* an indication of the density of the ratio of attenuation coefficients near a given value. Actually, there are eight modes having a_{xh}/a_{1h} ratios in the range 2.5 to 10, 19 modes in the range 10 to 20, and 15 modes greater than 20.

Experimental work reported elsewhere shows that significant conversion takes place between TE₀₁ and the TE₁₁, TE₂₁, TE₃₁ and TM₁₁ modes. There is some likelihood that conversion to TE₁₂ takes place, but the magnitude has not been measured. The experience gained by measurement, therefore, shows that most typical conversions occur between TE₀₁ and modes having attenuation ratios a_{xh}/a_{1h} in the range 2.5 to 12.

The absolute magnitudes of the conversion coefficients a_{1x} have in some cases been measured directly, and may also be inferred from measurements of total signal attenuation on the 500 ft. experimental line and separate knowledge of the heat-loss values; the inference is that a_{1x} must fall in the range 0.1 to 1.0 a_{1h} for the particular line studied.

REFERENCES

1. S. E. Miller and A. C. Beck, Low-Loss Waveguide Transmission, Proc. I. R. E., **41**, pp. 348-358, March, 1953.
2. S. E. Miller, Coupled-Wave Theory and Waveguide Applications, B. S. T. J., **33**, pp. 661-720, May, 1954.
3. M. Aronoff, Radial Probe Measurements of Mode Conversion in Large Round Waveguide with TE₀₁ Mode Excitation, presented orally at the March, 1951, I.R.E. National Convention.

4. M. Jouguet, Effects of the Curvature on the Propagation of Electromagnetic Waves in Guides of Circular Cross Section Cables and Transmission (Paris), **1**, No. 2, pp. 133-153, July, 1947.
5. M. Jouguet, Wave Propagation in Nearly Circular Waveguides: Transmission-Over-Bends Devices for H_0 Waves, Cables and Trans (Paris) **2**, No. 4, pp. 257-284, Oct., 1948.
6. W. J. Albersheim, Propagation of TE_{01} Waves in Curved Waveguides, B. S. T. J., **28**, pp. 1-32, Jan., 1949.
7. S. O. Rice, unpublished work.
8. S. E. Miller, Notes on Methods of Transmitting the Circular Electric Wave Around Bends, Proc. I.R.E., **40**, pp. 1104-1113, Sept., 1952.
9. A. C. Beck and R. W. Dawson, Conductivity Measurements at Microwave Frequencies, Proc. I.R.E., **38**, pp. 1181-1189, Oct., 1950.
10. J. B. Little, Amplification at 6-mm Wavelength, Bell Labs. Record, Jan., 1951.
11. S. Millman, A Spatial Harmonic Travelling-Wave Amplifier for Six-Millimeters Wavelength, Proc. I.R.E., **39**, pp. 1035-1043, Sept., 1951.
12. R. Kompfner, Backward-Wave Oscillator, Bell Labs. Record, Aug., 1953. R. Kompfner and N. T. Williams, Backward Wave Tube, Proc. I.R.E., **41**, pp. 1602, Nov., 1953.
13. A. Karp, Paper presented orally at the June, 1951, Conference on Electron Tube Research.
14. A. G. Fox, New Guided-Wave Techniques for the Millimeter-wave Range, presented orally at the March, 1952, I.R.E. National Convention.
15. J. R. Pierce and W. G. Shepherd, Reflex Oscillators, B. S. T. J., **26**, pp. 460-681, July, 1947.
16. G. D. Sims, The Influence of Bends and Ellipticity on the Attenuation and Propagation Characteristics of the H_{01} Circular Waveguide Mode, Proc. Institution of Electrical Engineers (London), **100**, Part IV, No. 5, pp. 25-34 Oct., 1953.
17. S. P. Morgan, Jr., Mode Conversion Losses in Transmission of Circular Electric Waves through Slightly Non-Cylindrical Guides, J. Appl. Phys., **21**, pp. 329-339, April, 1950.

A Governor for Telephone Dials

Principles of Design

By W. PFERD

(Manuscript received July 29, 1954)

This is a report on the development of a new type of governor for regulating the speed of rotary dials. The paper includes derivation of the equations of motion which determine the theoretical speed of the governor during dial run-down and an analysis of the operating characteristics of the governor as influenced by varying friction and input torque. Experimental verification of the relations is presented. A theoretical analysis which explains "governor chatter" or positional instability for friction-centrifugal governors is also given.

INTRODUCTION

Machine switching telephone systems depend on the telephone dial for originating information used in completing a call. During run-down, the dial originates current pulses which operate step-by-step switching equipment or are registered for use in common-control panel or crossbar systems. For reliable functioning of dial pulse controlled switching equipment, the pulses must be closely controlled in frequency and form. Since the pulses are produced during run-down of the dial after release by the customer, the run-down speed must be constant. Friction-centrifugal governors are commonly used to provide this required control of speed.

If the pulses reaching the central office were exactly like those generated by the dial, the designers of dials and central office switching apparatus and circuits would find themselves far less restricted. Unfortunately the dial pulses are distorted by the electrical characteristics of the customer's loop. To compensate for this distortion and insure accurate registration of the pulses at the central office, the dial and central office equipment must be designed to operate to close limits of performance. The designs must also be such that there is negligible change of

adjustment resulting from use, time or weather conditions, which might affect the ability to accurately send or receive dial pulses.

To achieve the accuracy of timing required of dial governors, it was recognized that a general theoretical analysis which defined speed of governors would be beneficial. The late C. R. Moore investigated this problem in the thirties, and derived from theoretical considerations, general equations of motion relating to governors. The relationships derived by the Moore analysis are extremely useful in that they can be used to indicate the influence of various design factors on the performance of a governor. This theory was applied in developing the new governor used in the 7-type dial of the 500-type telephone set* and will be presented in this paper. To better demonstrate the operating characteristics of the new governor, it will be compared with a previous governor which was used in an older type dial. Photographs of the new governor as it is assembled in a dial are shown in Figs. 1A and 1B.



FIG. 1A — Front view of 7-type dial.

* Inglis, A. H., and Tuffnell, W. L., An Improved Telephone Set, B.S.T.J., 30, pp. 239-270, April, 1951.

DIAL AND GOVERNOR OPERATION

In dialing, the fingerwheel is rotated through an angle proportional to the number being dialed and then released. Energy stored in the motor spring, Fig. 2, causes the mechanism to return to the start position. For each 30° rotation of the fingerwheel during run-down, the intermediate gear rotates one-half revolution and the cam pinion and pulsing cam rotate one full revolution. Once the pulsing pawl is in position, each revolution of the pulsing cam in the run-down direction results in a pulse being placed on the telephone loop. The intermediate gear also meshes with a governor pinion which is coupled to the governor shaft and governor through a spring clutch.* This clutch decouples the governor from the fingerwheel on windup to reduce the windup torque. On run-down the governor rotates two full revolutions for each 30° rotation of the fingerwheel.

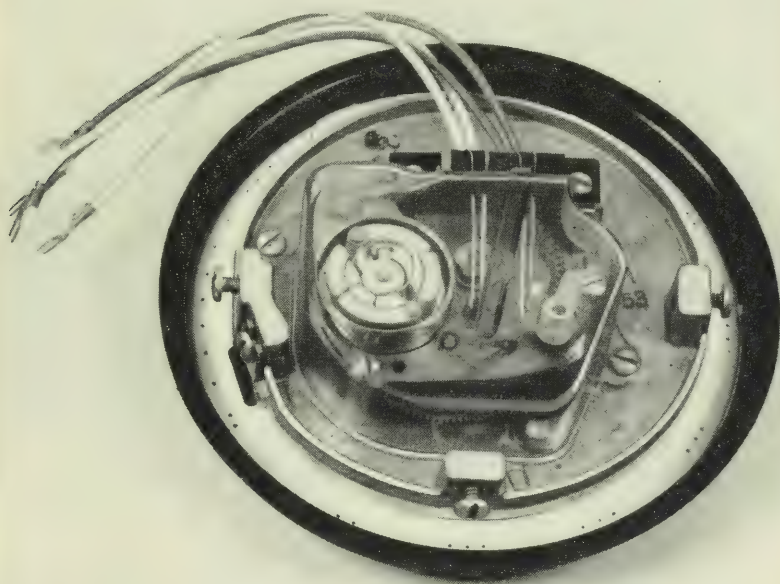


FIG. 1B — Rear view of 7-type dial. Left, speed governor, center, off-normal contact and right, pulsing mechanism.

* Wiebusch, C. F., The Spring Clutch, J. Appl. Mech., Sept., 1939.

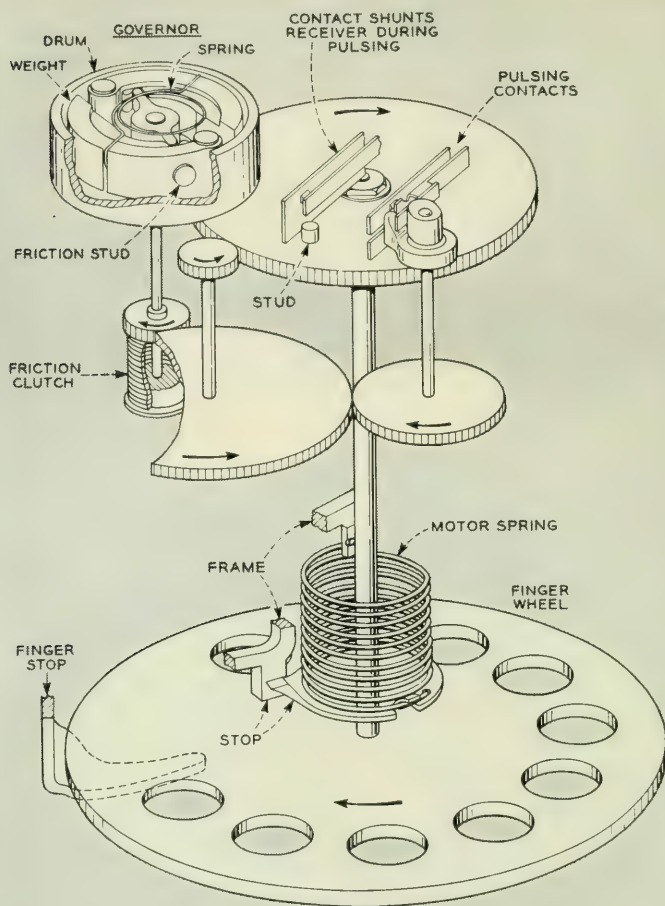


FIG. 2 — Simplified diagram of 7-type dial mechanism; governor appears top left.

As shown in Fig. 3, the weights of the new governor are free to pivot at the ends of the fly-bar which, in turn, is allowed to rotate with respect to the shaft. Rotation is imparted to the system by the drive-bar which presses against each weight at a specific point. As the mechanism begins to rotate during dial run-down, the two weights are caused by centrifugal force to move outward against the tension of a spring. Movement of the

weights about their pivots continues until the friction studs touch the case. At this instant governing begins, and controls the dial speed until the end of run-down. The speed attained by the governor will be dependent on the friction between the studs and case, the magnitude of the driving torque, and the tension to which the spring of the governor is adjusted.

In the schematic of the new governor, Fig. 4, the driving force is designated as F . By applying this driving force between the weight pivot and the center of the governor, the mechanism behaves during operation as a true friction-centrifugal governor and also as a brake. This configuration results in a gain in the ability of the governor to resist the increase in speed which normally results from an increase in the applied rotational force. The drive-bar force, F , and the torque produced by the stud-to-case force about the pivot assists the centrifugal force in pressing the friction studs against the case.

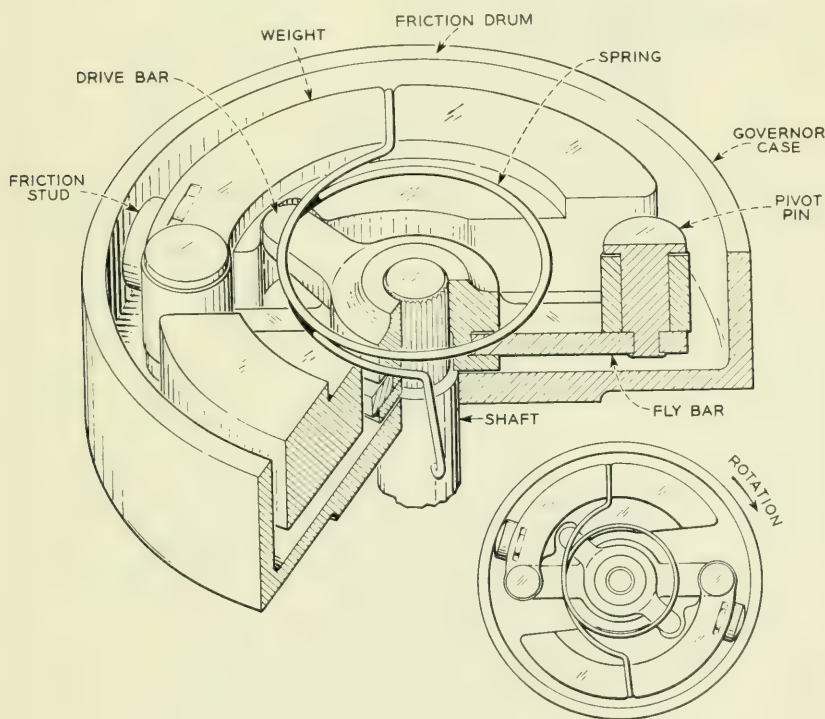


FIG. 3 — The 7-type dial drive-bar governor.

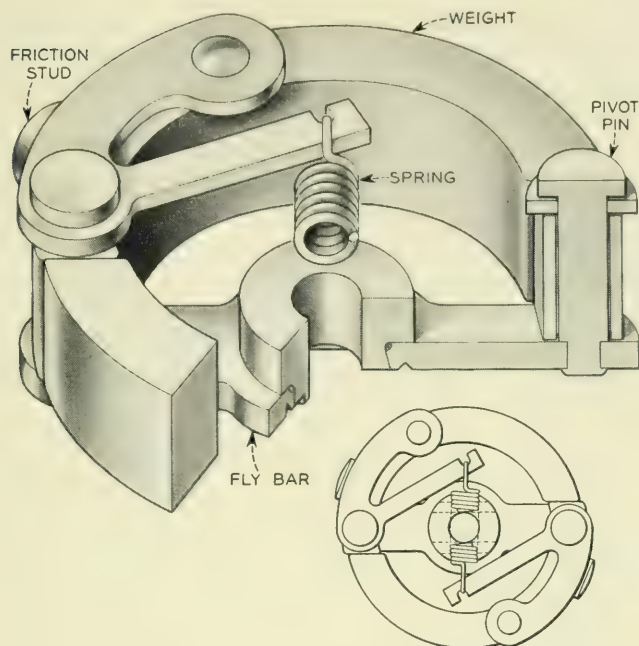


Fig. 5 — Fly-bar type governor.

friction about the pivots aids the centrifugal force in pressing the studs against the case.

EQUATIONS OF MOTION

In deriving the general equations of motion for a governor, two assumptions are made concerning the action. During the interval of time that the governor is approaching the critical velocity when contact of the friction surfaces first occurs, the motion is assumed to be that of a simple fly-wheel, constantly accelerating. The angular velocity of the governor during the time from rest to the critical velocity, ω_0 , is then given by

$$\omega = \frac{Gt'}{I} \quad (1)$$

where G = applied torque

t' = time from start of motion

I = moment of inertia of the governor assembly about the shaft

After stud-to-case contact occurs, it is assumed that there is no further pivoting of the weights or fly-bar, and the assembly rotates as a rigid body. During this time the general equation for angular velocity is

$$\omega = q \tanh \left(\frac{ht}{q} + \ln \sqrt{A} \right) \quad (2)$$

where q = regulated or final angular velocity

h = design constant

A = design and adjustment constant

t = time measured from the moment of initial braking

The derivation of this general speed equation as it applies to the new 7-type dial governor is given in Appendix I. For this drive-bar type governor the following terms apply:

$$h = \frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{I_0(d - \mu c)} \quad (3)$$

$$g = \frac{M\mu}{I_0(d - \mu c)} \quad (4)$$

$$A = \frac{q + \omega_0}{q - \omega_0} \quad (5)$$

$$q = \sqrt{\frac{h}{g}} = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu}} \quad (6)$$

The derivation of the theoretical equation for speeds in excess of the critical speed for the comparative fly-bar design results in the following relationships:

$$\omega = q \tanh \left(\frac{ht}{q} + \ln \sqrt{A} \right)$$

where

$$h = \frac{G(d - \mu c) + M\mu\omega_0^2}{I_0(d - \mu c)} \quad (7)$$

$$g = \frac{M\mu}{I_0(d - \mu c)} \quad (8)$$

$$q = \sqrt{\frac{h}{g}} = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2}{M\mu}} \quad (9)$$

The form of the general speed equation is the same for all types of

friction-centrifugal governors but each particular governor will have different terms in the values for g , h and q . The theoretical equation of motion can be used to calculate the speed of the dial at any time, t , after the critical velocity is reached or the time required to reach any given speed once the governor studs touch the case. The equation shows that for large values of t , ω approaches q , so that steady state speed is given by the value of q for each type governor, and is in terms of the operating values and design constants of the mechanism.

For the drive-bar governor the steady state speed equation is

$$\omega = q = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu}} \quad (10)$$

THEORETICAL SPEED-TIME CURVES

Drive-Bar Governor

The design constants and physical data given in Table I apply to the drive-bar governor, and were used to calculate the theoretical speed

TABLE I — REFER TO FIGS. 3 AND 4

$d = 0.390$ cm	$I_0 = 7.40$ gm cm ² (experimental)
$c = 0.361$ cm	$G = 7,500$ dyne-cm (steady-state)*
$r = 1.180$ cm	13,500 dyne-cm (initial)
$r_0 = 0.635$ cm	$\mu = 0.25$ (assumed)
$b = 0.920$ cm	$m = 3.9$ gms
$e = 0.498$ cm	$K = r_0/r = 0.538$
$j = 0.236$ cm	$M = 2mr^2bk = 5.38$

* Appendix II — Governor Input Torque.

versus time curve shown in Fig. 6. The dial governor is initially adjusted so that signaling is at the rate of 10.0 pulses per second, requiring a steady state governor shaft velocity of 125.6 radians per second. This steady state velocity was used to determine the critical velocity ω_0 by substituting the values in Table I in equation (10):

$$\omega = q = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu}}$$

$$\omega_0 = 121.8 \text{ radians/second}$$

and from equations (3), (4), (5), and (6)

$$g = 0.606 \quad h = 9,520 \quad h/q = 75.9 \quad A = 72.7.$$

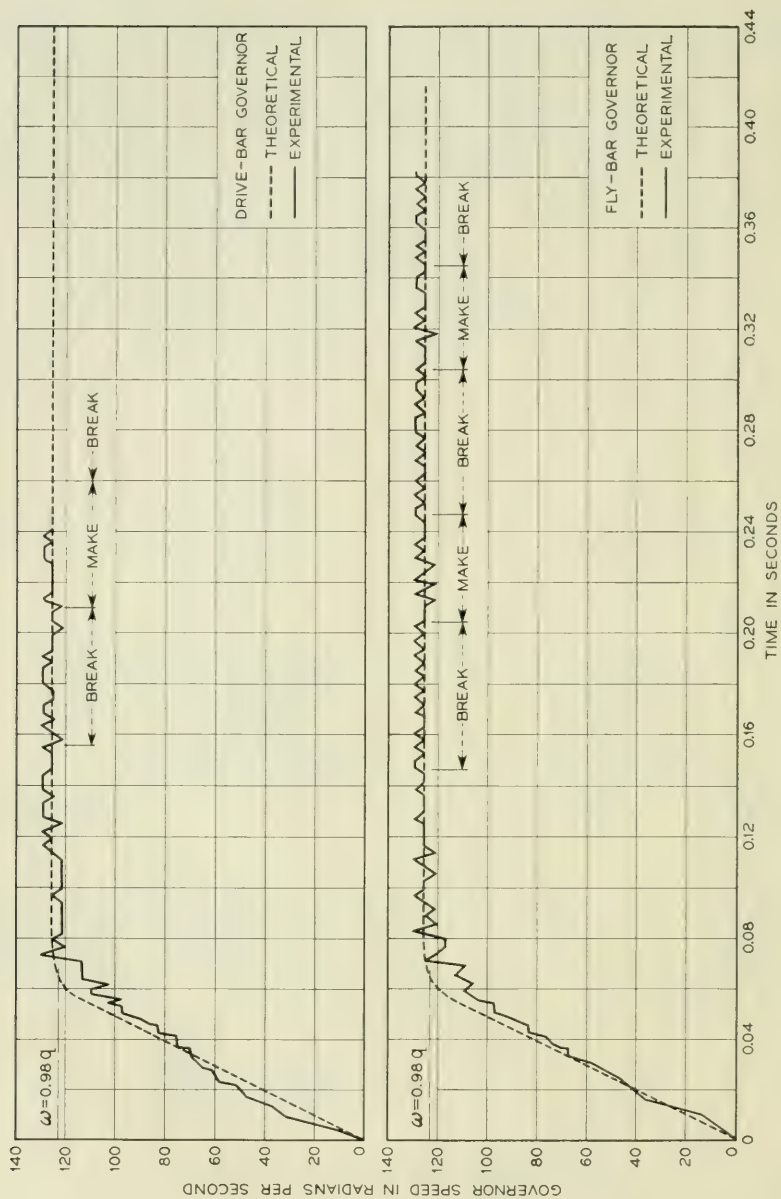


Fig. 6 — Speed curves for the drive-bar and fly-bar type governors.

TABLE II

t	$75.9t + 2.142$	$\omega = 125.6 \tanh (75.9t + 2.142)$
0.000	2.142	121.8
0.004	2.446	123.7
0.008	2.750	124.6
0.012	3.054	125.1
0.016	3.358	125.3
0.020	3.662	125.5

Substituting these values in the general speed equation,

$$\omega = q \tanh \left(\frac{ht}{q} + \ln \sqrt{A} \right)$$

gives the velocity of the drive bar governor at any time t measured from the moment the governor reaches the critical velocity, i.e., when the friction studs first touch the inside of the governor case. (Table II.)

For governor speeds from start of rotation up to the critical velocity, it is assumed that the system rotates as a simple fly-wheel, therefore from equation (1)

$$t' = \frac{\omega_0 I_0}{G_I} = \frac{121.8(7.4)}{13,500} = 0.0668 \text{ seconds}$$

This time of 0.0668 seconds determines the slope of the straight line portion of the theoretical speed-time curve shown on Fig. 6.

Fly-Bar Governor

The data given in Table III applies to the fly-bar governor shown schematically in Fig. 7. Substituting the values given in this table in the steady state speed equation for the fly-bar governor,

$$\omega = q = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2}{M\mu}}$$

$$\omega_0 = 118.5 \text{ radians/sec.}$$

and from equations (7), (8) and (9):

$$g = 0.609 \quad h = 9,560 \quad h/q = 76.4 \quad A = 36.35$$

Substituting these values in the general speed equation gives the velocity of the fly-bar governor at any time (t) measured from the moment braking begins, Table IV. For this particular fly-bar design, the

time required to reach the critical velocity is, from equation (1),

$$t' = \frac{\omega_0 I_0}{G_I} = \frac{118.5(7.36)}{13,500} = 0.0646 \text{ seconds}$$

EXPERIMENTAL SPEED-TIME CURVES — NORMAL TORQUE

Experimental velocity versus time curves were obtained for the fly-bar and drive-bar governors constructed to the specifications listed in Tables I and III. Data used in determining the true speed versus time picture for the experimental governors was taken from photographic traces exposed on a recording oscillograph. A thin disc having 36 radial slots spaced every 10° was fastened to the end of the governor shaft. Light detected through the slots of the moving disc by the element of a photo tube was used to deflect one of the strings of the oscillograph. The trace

TABLE III — REFER TO FIGS. 5 AND 7

$d = 0.390 \text{ cm}$	$G = 7,500 \text{ (steady state)}$
$c = 0.361 \text{ cm}$	$13,500 \text{ (initial)}$
$r = 1.18 \text{ cm}$	$\mu = 0.25 \text{ (assumed)}$
$r_0 = 0.635 \text{ cm}$	$m = 3.9 \text{ gm}$
$b = 0.920 \text{ cm}$	$K = r_0/r = 0.538$
$I_0 = 7.36 \text{ gm cm}^2$	$M = 2mr^2bK = 5.38$

of this string appeared on the photographic paper as a distorted sine wave. The distance between two successive wave peaks represented 10° of rotation of the governor. By noting on the trace the time between peaks, it was possible to determine the average velocity of the governor at 10° intervals after release of the finger-wheel, or start of rotation of the governor mechanism. The experimental speed curves for the drive-bar and fly-bar governors are plotted on Fig. 6 along with the theoretical speed curves.

It was assumed in the theoretical analysis that while accelerating up to the critical velocity, the governor assembly rotates as a simple fly-wheel. This requires that the velocity increase linearly. The theoretical and experimental velocity curves for both type governors during the initial accelerating period show the fly-wheel assumption to be justified. The slope of the velocity curve, or rate of acceleration is generally constant.

For that portion of the theoretical and experimental curves which show speeds from the critical velocity to 98 per cent of rated speed, agreement is not too clearly defined. The theoretical curve is naturally smooth in shape. An oscillating type characteristic appears in the experimental

speed curves of both types of governors. This probably results from the shock and grabbing when the friction studs first touch the case and continues until the forces tending to move the weights outward increase to a value sufficient to hold them against the case for governing.

That part of the experimental curve, during which governing actually occurs at rated speed is relatively smooth through full run-down. Both the fly-bar and new drive-bar governors exhibit excellent speed regulation. The waves present on the trace do not necessarily indicate hunting or vibration since the variations in speed which appear are actually smaller in magnitude than the degree of accuracy present in measuring the experimental photographic trace.

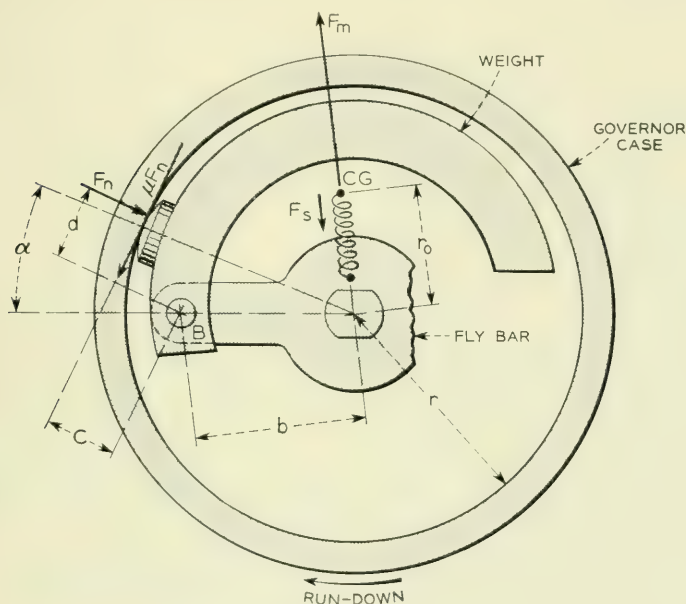


Fig. 7 — Schematic of fly-bar governor.

- F_n — Normal force of case acting on studs
- F_s — Force exerted by spring when studs touch case
- F_m — Centrifugal force acting at center of gravity of each weight
- μ — Coefficient of friction
- I_0 — Moment of inertia of the governor about center shaft
- ω — Angular velocity of governor
- ω_0 — Critical angular velocity at which studs just touch the case
- m — Mass of each weight
- r_0 — Radius to the center of gravity of each weight
- r — Radius of governor case
- α — Stud angle

Neg. Rotation — Run down of governor

THEORETICAL OPERATING CHARACTERISTICS

The solution of the equation of motion for any governor is based on specific design constants and certain assumed and estimated values. Such dimensions as the governor case inside diameter r , and distance from the shaft center to the weight pivot b , are two examples of design constants. These constants establish the arrangement of the various component parts of the mechanisms and are, therefore, subject to practical manufacturing and space considerations as well as considerations from a speed regulation standpoint. The design constants are in effect static considerations. They are not subject to appreciable variation once established, and on any particular governor do not vary significantly over the life of the governor.

TABLE IV

t	$76.4t + 1.798$	$\omega = 125.6 \tanh (76.4t + 1.798)$
0.000	1.798	118.5
0.004	2.103	121.9
0.008	2.408	123.6
0.012	2.714	124.5
0.016	3.010	125.0
0.020	3.325	125.3
0.024	3.631	125.4

During actual operation there are two factors which directly affect the degree of speed control afforded by a governor; i.e., the coefficient of friction which exists between the governor case and friction studs, and the value of input torque to the governor shaft. Design control over these factors is present, but to a lesser degree than for the design constants mentioned previously. These factors are considered fixed in arriving at a given design but actually vary from governor to governor and over the life of a governor. It is therefore necessary to consider carefully the effect of variations in friction and torque if close regulation of speed is required.

The input torque to the governor shaft will vary because of the dimensional variations of the motor springs, the tolerance permitted for the driving torque at full windup of the dial, dial friction and the variation in pulsing spring forces. These variables appear at the governor as a range of input torques during run-down of the mechanism. For the motor spring used in the 7-type dial, input torque referred to the governor shaft decreases during run-down on the average from 20,000 dyne-cm to 13,000 dyne-cm. Torque required to overcome bearing and gear fric-

tion and the loads imposed by the pulsing mechanism result in an average torque of 7,500 dyne-cm at the governor. Over the life of a dial this input torque at the governor will vary as the dial efficiency varies. Initially the dial mechanism is lubricated and the bearings and gears turn freely. With time and continued operation, the accumulation of dirt and wear products affect the dial so that more torque is needed to drive the moving parts. This causes a decrease in the remainder torque going to the governor.

Another aspect of torque requiring consideration is that resulting from forcing of the finger-wheel during run-down. This action can produce torque values at the governor of the order of 110,000 dyne-cm or a torque of approximately fifteen times that which appears at the governor during normal operation.

The second factor, which can vary during dial life, is the value of the stud-to-case coefficient of friction, μ . Both the drive-bar and fly-bar governors have studs of Ebonite compounded with 40 per cent by weight of hard rubber dust and cases of ASTM B16 brass. Actual service tests show the satisfactory wearing ability of these materials.

Each governor is initially adjusted for speed by changing the tension of the governor spring. At the time this adjustment is made, a particular friction condition exists between the governor studs and case. With time or continued operation there is always the possibility of a change occurring in this friction value. Such factors as very high humidity, lubrication products traveling to the stud operating region, or the accumulation of foreign-material or wear particles may produce different values of friction and hence result in variation in governor and dial speed from the initially adjusted value.

The range of coefficient of friction values μ expected for rubber on brass is from 0.05 to 0.35. These are the extreme conditions produced by oil in the governor case for the 0.05 value and very low unit pressure on a scored brass surface for the 0.35 value. For this study a representative figure for the average stud-to-case friction value was taken to be 0.25.

The problem of variation in steady state governor speed with changes in the coefficient of friction and input torque can be analyzed by considering the derivatives of speed with respect to these values. This is done by operating on the equations for terminal speed.

Speed with Respect to Coefficient of Friction

For the drive-bar governor, the partial derivative of speed, with respect to coefficient of friction, is as follows:

$$\omega = q = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu}}$$

Taking the derivative of ω with respect to μ gives

$$\frac{\partial \omega}{\partial \mu} = -\frac{Gd}{2\omega M\mu^2} \quad (11)$$

where $M = 2mrr_0b$. For the fly-bar governor, $\partial\omega/\partial\mu$, is as follows from

$$\omega = q = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2}{M\mu}} \quad (12)$$

$$\frac{\partial \omega}{\partial \mu} = -\frac{Gd}{2\omega M\mu^2}$$

Since the equations are identical for both governors the same considerations exist in holding to a minimum the change in speed caused by a change in the coefficient of friction.

For optimum speed regulation, the partial derivative, $\partial\omega/\partial\mu$, should be a minimum. Small values of $\partial\omega/\partial\mu$ can be obtained by operating on the design constants, controlling the value of μ , or having high governor speeds. Specifically, the design constants m , r_0 , r and b should all be large. Inspection of the drive-bar governor schematic, Fig. 4, shows that there are physical limitations to the arbitrary enlargement of these values. Space, manufacturability, and cost of materials must be considered in establishing these terms. In selecting materials for fixing the coefficient of friction value the wearing quality of the materials to be used must be of first consideration. A high governor speed is advantageous but must be weighed against the primary disadvantage of high inertia loads for the entire mechanism.

The terms G and d in the numerator of equations (11) and (12) indicate that low input torque and a small value for d would be desirable. For G , one must consider the anticipated change in dial efficiency and variation in torque produced by the pulsing mechanism plus the torque necessary for the governor to maintain adequate speed control.

As shown in the drive-bar governor schematic, Fig. 4, d is the distance from the weight pivot to the normal of the point of contact between the rubber stud and governor case. Its magnitude is controlled by the stud angle α , and the distance between the stud and case when the weights are in the closed position.

As indicated, d should be as small as possible for best regulation with friction change. This requirement imposes a difficult design problem because as d decreases, the stud and bearing hole in the weight approach

each other. A minimum d is therefore fixed by interference of the parts themselves. Inspection of the governor mechanism also shows that even if the interference problem were not limiting, the weight turning angle imposes a further restriction on the d value. The life of a governor is considered to end when the friction material is worn to the point of allowing the weight to touch the inside of the case. Because of this initially large weight motion, little material is provided for wearing and hence the possible life of the mechanism is reduced. In the design of the drive-bar governor, the stud angle, α , was made as small as manufacturing techniques would permit. The stud angle is 22° from the weight pivot with a corresponding $d = 0.390$ cm.

Since both governors under study were designed to operate in the case of the same dial the terms shown in Tables I and III which apply in the friction equations (11 and 12) are identical. This identity of terms indicates that both type governors should exhibit identical frictional characteristics. Fig. 8 represents a plot of the $\partial\omega/\partial\mu$ for various governor input torque values. One set of curves applies. The range of torque values covered by the curves is from zero to the forcing condition at fifteen times normal motor spring torque. Curves for coefficient of friction values, 0.05, 0.10, 0.20, 0.35, are shown as encompassing the extremes in operating range.

Speed with Respect to Input Torque

For the drive-bar governor, the partial derivative of speed with respect to torque is determined as follows from the steady state speed equation.

$$\omega = q = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu}}$$

Taking the derivative of ω with respect to G gives:

$$\frac{\partial\omega}{\partial G} = \frac{1}{2} \left[\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu} \right]^{-1/2} \frac{(d - \mu c - \mu r j G/e)}{M\mu}$$

and

$$\frac{\partial\omega}{\partial G} = \frac{1}{2} \frac{(d - \mu c - \mu r j/e)}{M\mu\omega} \quad (13)$$

where

$$M = 2 m r r_0 b$$

For the fly-bar governor the derivative of ω with respect to G can be

developed similarly to give

$$\frac{\partial \omega}{\partial G} = \frac{1}{2} \frac{(d - \mu c)}{M \mu \omega} \quad (14)$$

Here also the design constants, m , r_0 , r and b , the coefficient of friction μ , and the governor speed ω , should all be high values. Minimum change in dial speed would then occur for a given change in the input torque.

Inspection of the equations for $\partial \omega / \partial G$, indicates that it is possible to have perfect torque regulation for at least one value of μ . For this limiting condition, $\partial \omega / \partial G$, would equal zero and equation (13), for the drive-bar governor, equates to

$$d - \mu c - \mu r j / e = 0 \quad (15)$$

and equation (14) for the fly-bar type governor

$$d - \mu c = 0 \quad (16)$$

If these equations were satisfied, there would be zero change in speed for a given change in input torque to the governor. As stated previously μ is predetermined and has in this design a maximum known value of 0.35. A margin of safety is considered by taking $\mu = 0.425$ for the limiting case and equation (15) becomes for the new drive-bar governor

$$d - 0.425 (c + r j / e) = 0 \quad (17)$$

and equation (16) for the fly-bar governor

$$d - 0.425 c = 0 \quad (18)$$

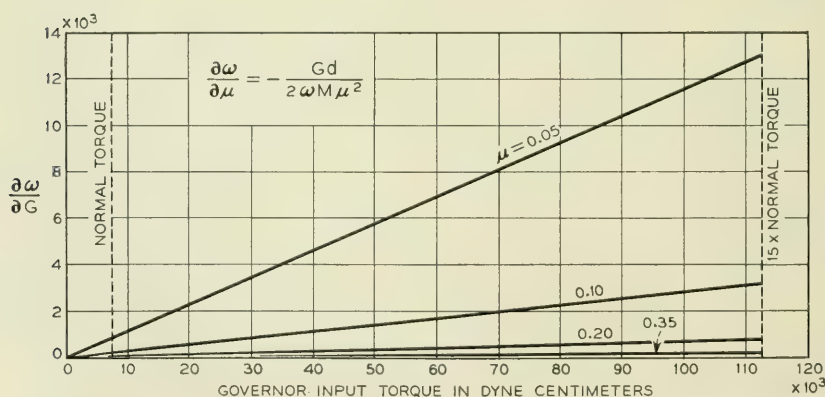


Fig. 8 — Derivative of governor speed with respect to coefficient of friction versus input torque to the governor.

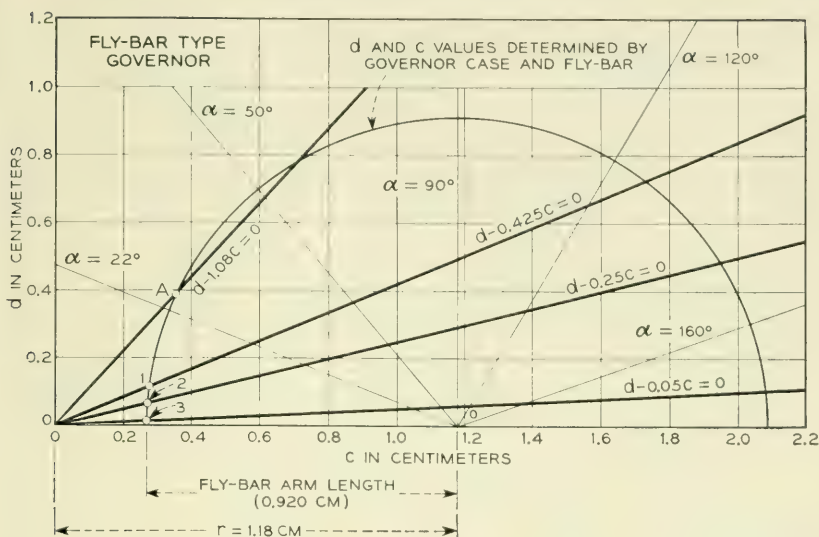


Fig. 9 — Design diagram for fly-bar type governor.

Comparison of these two equations shows a very important difference in the term multiplied by $\mu = 0.425$. For the drive-bar governor, there are four variables which can be operated on to satisfy the equation; i.e., c , r , j and e . For the fly-bar governor only c is available. The importance of these additional terms can be realized when one considers that the value for c results from our choice of d in making $\partial\omega/\partial\mu$ a minimum. For both type governors c is equal to 0.361 cm. Substitution of the d and c values in the limiting equation, (18), for the fly-bar governor does not lead to a solution when $\mu = 0.425$. Solving for this limiting μ in the fly-bar governor gives

$$(0.390) - (0.361) \mu = 0$$

$$\mu = 1.08$$

This value of μ is far beyond that encountered in actual governor operation and, in effect, represents useless margin. This is graphically shown in Fig. 9 where all d and c values which conform to the geometry of the fly-bar governor mechanism are plotted as a design diagram. The three straight lines radiating from the origin represent plots of the limiting equation for $\mu = 0.425, 0.25$ and 0.05 . The intersection of these lines with the d and c semicircle, noted at points 1, 2 and 3, give the particular angle at which the stud should be located for optimum regula-

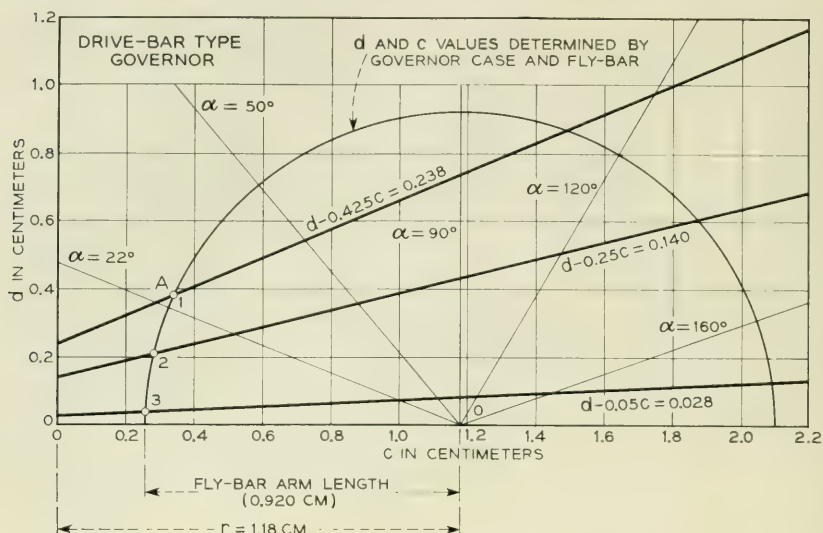


Fig. 10 — Design diagram for drive-bar type governor.

tion for these particular friction values. The point A on the d and c circle denotes the d and c values which result from the stud being at $\alpha = 22^\circ$. The off-set appears because stud-to-case contact is not made on the center line of the stud. The points 1, 2 and 3 are all below the point, A, established by the minimum permissible angle, $\alpha = 22^\circ$. This indicates that the fly-bar governor has its $\partial\omega/\partial G$ equal to zero at some coefficient of friction value higher than $\mu = 0.425$. As previously determined in this value is $\mu = 1.08$.

Fig. 10 represents the design diagram for the new drive-bar governor. Here again, the large semi-circle is a plot of all d and c values which conform to the geometry of this governor mechanism. Point A is determined by the stud angle $\alpha = 22^\circ$. The straight lines represent the limiting equations for $\mu = 0.425$, 0.25 and 0.05 and are shown intersecting the d and c circle at 1, 2 and 3 respectively. For this particular governor, the line representing $\mu = 0.425$,

$$d - 0.425(c + rj/e) = 0$$

intersects the d and c circle at the point A. This is possible by making the term $(\mu rj/e)$ equal to 0.238. The intersection of this curve at A indicates that there will be zero change in speed for a given change in input torque to the governor when the stud-to-case coefficient of friction value is 0.425.

The additional terms, r , j and c in the limiting equation make it possible to design the governor for optimum regulation for any particular value of μ desired. Since the term r , the case inside radius, is controlled by space requirements, the terms j and c assume added importance. They are determined by the point at which the drive-bar arms act against the weights and can be made any value required to meet the design objective.

Figure 11 is a plot of $\partial\omega/\partial G$ at various values of μ for the drive-bar and fly-bar governors specified in Tables I and III. It indicates that for any coefficient of friction the new drive-bar governor should exhibit less change for a change in input torque than is possible with the fly-bar governor.

EXPERIMENTAL DATA

To substantiate the theoretical conclusions drawn from the analysis of the steady state speed equations, experimental data were compiled from a number of models of each type of governor. Drive-bar and fly-bar governors, made to the specifications listed in Tables I and III were investigated to determine their response to changes in input torque and changes in the stud-to-case coefficient of friction value, μ . The tests were conducted on 7-type dials manufactured by the Western Electric Company as standard product.

Dial Speed Versus Coefficient of Friction

The theoretical analysis of the equation

$$\frac{\partial\omega}{\partial\mu} = -\frac{Gd}{2\omega M\mu^2}$$

indicates that the two types of governors should exhibit identical frictional characteristics. Fig. 12 represents the theoretical plot of dial speed in pulses per second versus coefficient of friction. The single curve satisfies both the fly-bar and drive-bar governors as specified in Tables I and III. This curve shows the change in speed of a dial initially adjusted to 10.0 pulses per second, operating at normal torque, as the coefficient of friction varies. It indicates that if there is a decrease in the value of μ from that which existed at the time of initial adjustment, there will be a corresponding increase in the dial speed. The curve is drawn with $\mu = 0.25$ as representing the normal stud-to-case condition and a normal governor input torque of 7,500 dyne-cm.

The experimental data were compiled for the following operating con-

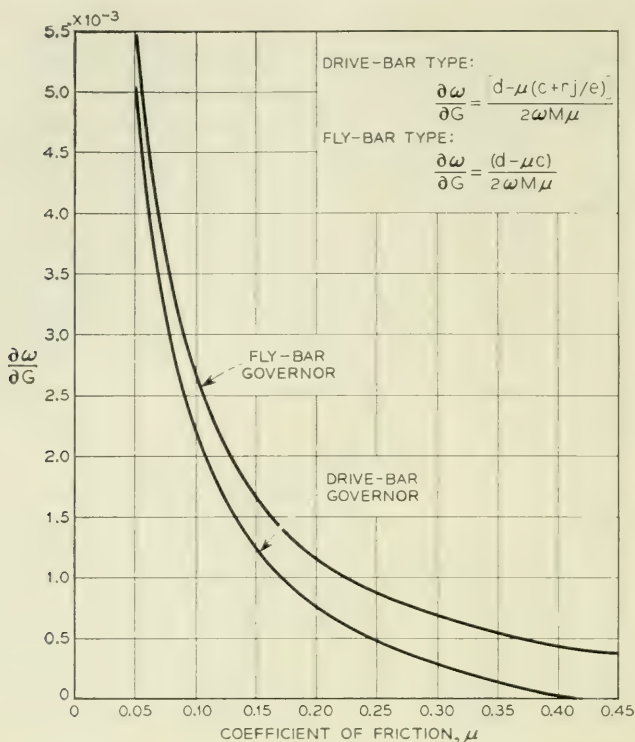


Fig. 11 — Derivative of governor speed with respect to governor input torque versus coefficient of friction.

ditions: as received, governor case cleaned with acetone, damp atmosphere, and SAE 10 petroleum base oil in the case. Each of the governors were initially adjusted to 10 pulses per second with the governor cases in the "as received" condition. The governors were then removed and the cases and governor studs were cleaned with acetone. The governors were then reassembled and the new speed recorded. During this procedure extreme care was exercised so as not to disturb the governor spring adjustment. Speed was next recorded for the damp atmosphere condition, and finally for the condition with one drop of SAE 10 oil in the governor case. For these last two conditions the governors were not removed from the dials.

The average speeds recorded for the four conditions are plotted on the theoretical speed curve of Fig. 12. The points were arbitrarily placed on the theoretical curve. No attempt was made to determine the exact coefficient of friction values corresponding to the four conditions. In this

respect, the speed attained with oil in the case shows a value of $\mu = 0.06$ which is very close to the $\mu = 0.05$ taken as the lower limit.

To produce a decrease in governor speed one must increase the value of μ . This is difficult to do on a controlled basis, since it is brought about by the progressive action of wear particles and foreign material scoring the surface of the brass case during the life of the governor. The theoretical analysis indicates that when the coefficient of friction increases to $\mu = 0.35$ the speed of the governor will decrease from 10.00 to 9.65 pulses per second. This assumes that there would be no speed change due to wear in any part of the governor. In practice of course the governor studs wear as the case is scored and, therefore, are made progressively shorter. For this condition the increased outward motion of the weight required for stud-to-case contact produces an increase in the spring force, F_s , acting on the weights. The speed change which results from increasing the spring force is in the direction to compensate for the loss in speed due to increase in coefficient of friction. Therefore, considering only stud friction and wear as effective in causing change in speed, generally the governor and dial speed should increase from its initially adjusted value during life.

It can be concluded from the experimental and theoretical evidence that there is the possibility of a speed change due to varying coefficient

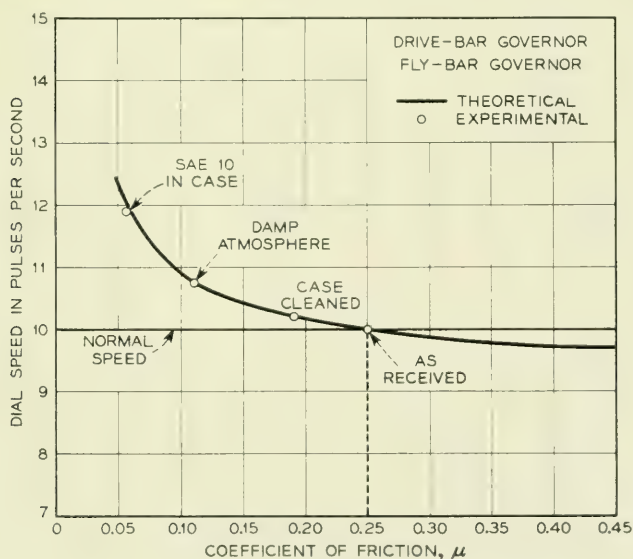


Fig. 12 — Dial speed versus coefficient of friction.

of friction to values between 9.65 and 12.0 pulses per second during life for dials initially adjusted to 10.00 pulses per second. An increase in speed may result from a decrease in stud friction due to the existence of lubricant or high humidity in the stud operating region or, the speed may decrease due to high friction.

The above changes represent the extremes in dial speed determined solely by reaction of the governor to change in stud friction. In practice it is anticipated that dials adjusted to 10 pulses per second initially can vary from 9 to 11 pulses per second during normal usage. A reduction in speed will occur as more torque is required to compensate for the increase in bearing friction caused by the accumulation of dirt and wear products during ordinary life. This additional bearing drag will cause a decrease in the torque available for governing and therefore the dial will be regulated at reduced speed. For extreme cases of wear and contamination, it is of course possible that the dial will stop altogether during run-down. Such cases are not controllable by the governor. They result from the expected attrition during extended life or unfavorable environment.

To guard against excessive increase in dial speed from the value at time of initial adjustment, precaution is taken during manufacture. As stated previously, lubricant traveling into the governor case after initial adjustment will cause a marked increase in dial speed. To avoid this sort of contamination, the governor case is washed after machining and swabbed with clean chamois prior to insertion of the governor onto the shaft. Care is also taken to see that no lubricant is placed in the governor case during lubrication of the shaft bearings. These practices assure that initially the friction surfaces are relatively free from contamination. The increase in dial speed up to the 11 pulses per second possible during dial life will result primarily as a result of stud wear, increase in efficiency of the mechanism, and operation during periods of high humidity.

Dial Speed Versus Governor Input Torque

To substantiate the theoretical conclusion that the drive-bar governor should exhibit better regulation due to changes in input torque, experimental data were compiled on the dials equipped with the two type governors. The results of this test are plotted on Fig. 13, along with theoretical forcing curves for both governors. A coefficient of friction value of 0.25 was assumed in arriving at the theoretical curves.

The dials were initially adjusted to 10.00 pulses per second by bending the governor spring to have proper tension. Loads of 1, 3 and 5 lb were

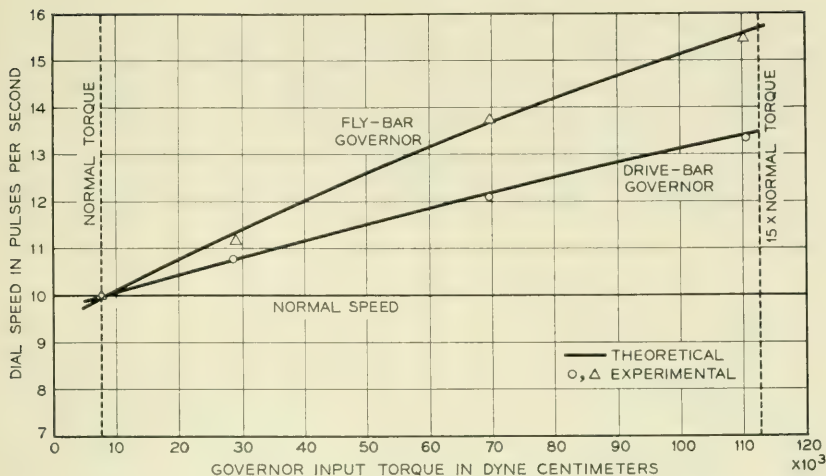


Fig. 13 — Dial speed versus governor input torque.

suspended from the fingerwheel at $\frac{5}{8}$ " radius and released. Average experimental speeds were recorded for the three forcing conditions and are noted in Fig. 13 for both the fly-bar and the new drive-bar governors. Good agreement, between the theoretical and experimental values, is evident. For a forcing condition of fifteen times normal motor spring torque, an average speed of 15.6 pulses per second is shown for the fly-bar governor while an average speed of 13.4 pulses per second is noted for the drive-bar type. Theoretically the speeds should be 15.3, and 13.2, respectively. This type agreement is also present for the 1 and 3 lb forcing conditions, and therefore, it may be concluded that for any input torque resulting from forcing the fingerwheel a dial equipped with a drive-bar governor will exhibit less speed increase than one having the fly-bar governor.

The theoretical analysis indicates that for the torque available during normal rundown, drive-bar governors as specified in Table I will decrease in speed from 10.00 to 9.80 pulses per second and fly-bar governors as specified in Table III will decrease in speed to 9.70 pulses per second. These theoretical speed changes were checked experimentally by recording on a rapid record oscillograph a trace of the make and break times of the pulsing contacts during rundown of the dial from digit zero. This information was used to determine the average dial speed in pulses per second for each sequence of make and break times. Actual loss in speed from the first to the ninth pulse for dials equipped with drive-bar gover-

nors averaged 0.24 pulses per second while dials with fly-bar governors decreased 0.33 pulses per second. The slight additional loss in speed noted experimentally was probably due to friction in the gear mechanism which is not considered by the theoretical analysis. However, since the differences recorded are quite small, one can conclude that the theoretical and experimental results are in good agreement even when concerned with small changes in torque experienced during normal rundown of a dial.

CHATTER IN GOVERNORS

It is not uncommon for governors with fine regulating ability to produce an objectionable chattering noise when operated near or at the vertical position. This chattering, while in most cases not particularly adverse from a regulation or wear point of view creates in the mind of the listener grave doubts as to the correctness of the design. In severe cases a sharp noise is heard during every half revolution of the governor shaft as each weight alternately leaves the case and strikes against the end of the other governor weight. During every revolution of the governor shaft, each weight is alternately supported as shown in the schematic, Fig. 4. At this instant, the gravity moment about *B* is a maximum, and along with the spring moment, opposes the centrifugal moment. If the gravity component, or effective mass of the weight, is sufficiently large, a new system is produced which has a critical velocity in excess of the regulated speed. Since the governor speed is continually regulated by the bottom weight at a speed lower than the new critical velocity, the top weight falls from contact with the case.

The magnitude of the gravity component is a function of the angle at which the governor operates and is a maximum when the dial is in the vertical position. As the operating plane of the dial decreases to the horizontal, the gravity effect decreases to zero. Chatter will not occur when the operating angle produces a gravity component smaller than the difference between the centrifugal force and the spring force.

Since the chattering effect is the result of a balance of forces on the governor, it is apparent that a relationship can be derived which will express the effect in terms of governor constants. This derivation is given in Appendix III and shows the chatter equation for a conventional fly-bar governor to be

$$m \leq \frac{G(d - \mu c)}{2r\mu\ell \sin \beta} \quad (19)$$

This expression must be satisfied if the governor is to operate free of chatter.

By substituting the constants for the fly-bar governor, in Table III, we have for this governor operating in the vertical plane

$$3.9(980) \leq \frac{7,500(0.390 - (0.25)(0.361))}{2(118)(0.25)(1.092)}$$

or

$$3,820 \text{ dyne-cm} \leq 2,790 \text{ dyne-cm.}$$

Since the equation is not satisfied instability should be present and governors of this fly-bar design do chatter loudly when operating in the vertical plane.

The chatter equation for the fly-bar governor indicates that by adjusting the design constants, one can eliminate the instability effect. This is true. A set of values could be used which would result in a fly-bar governor which operates free of chatter. Unfortunately such a governor would also have reduced ability to govern. The relationship between chatter and governing is explained as follows.

The equations which define changes in governor speed with respect to changes in friction and torque for the fly-bar governor

$$\frac{\partial \omega}{\partial \mu} = -\frac{Gd}{2\omega M\mu^2}$$

and

$$\frac{\partial \omega}{\partial G} = \frac{1}{2} \frac{(d - \mu c)}{M\mu\omega}$$

and the chatter equation, (19), show that operation without chatter and good speed regulation are totally incompatible. Those terms in the equations which should be small for good speed regulation; i.e., torque G and stud location d and c , should be large to avoid chattering of the governor. Those terms which should be large for good speed regulation; i.e., case radius r , friction μ , and the distance from the pivot to the center of gravity l , must be small for no chatter. As the theoretical analysis indicates there is no term in the fly-bar governor chatter equation which can be operated on to eliminate chatter without impairing regulation of speed.

A similar analysis, given in Appendix IV, shows the chatter equation for the new drive-bar governor to be

$$m \leq \frac{G(d - \mu c - \mu rj/e)}{2r\mu\ell \sin \beta} \quad (20)$$

This expression must be satisfied if there is to be no chattering during operation. A comparison of this equation with that determined for the fly-bar governor, equation (19), shows an additional term $(\mu r j/e)$. Substitution of the design constants for the drive-bar governor given in Table I leads to the following.

$$3.9(980) \leq \frac{7,500(0.390 - (0.25)(0.361) - (0.25)(1.18)(0.236)/(0.498))}{2(1.18)(0.25)(1.092)}$$

or

$$3,820 \text{ dyne-cm} \leq 1,862 \text{ dyne-cm.}$$

This indicates that instability should be present in this governor, and that the additional term $(\mu r j/e)$ causes a greater difference between the mass term and the torque term, than for the conventional fly-bar governor. This is to be expected, since for the drive-bar governor also, adequate speed regulation, and a design which has no chatter, are totally incompatible. The sensitivity of speed to torque change and changes in coefficient of friction.

$$\frac{\partial \omega}{\partial G} = \frac{1}{2} \frac{(d - \mu c - \mu r j/e)}{M \omega \mu}$$

and

$$\frac{\partial \omega}{\partial \mu} = -\frac{Gd}{2M\omega\mu}$$

were made as small as possible for best regulation and this results in a small value to oppose chatter.

This chatter analysis indicates that a new approach in design is needed in order to provide a governor which will operate without excessive noise and still regulate speed as required for use in the 7-type dial. This is found in a design which prohibits rapid movement of the governor weights during the unstable period.

Referring to the assembly drawing of the drive-bar governor, Fig. 3, which shows the governor in the rest position, one can see that each governor weight rests on the end of one of the arms of the drive-bar. By supporting the weights in this manner the following two beneficial effects are achieved. One, during operation in new assemblies, the weights move outward to touch the case only a nominal distance of 0.007". This small allowable motion in the drive-bar governor results in a low velocity of the weight at closure and hence, less impact noise. Two, because the drive-bar presses to rotate the governor weights, impact occurs as the

unstable weight skids against its arm of the drive-bar. This introduces friction damping to still further reduce the noise on closure of the weight. Some additional damping is provided by making the drive-bar and weight of powdered metal rather than wrought material. These features, which result from the particular physical arrangements of the component parts make possible the relatively quiet operation of the drive-bar governor. Experience with dials equipped with drive-bar governors indicates that they are effective since the noise due to chattering has been satisfactorily reduced so as to not be objectionable.

SUMMARY

This study has carried forward the work of C. R. Moore by presenting the derivation of theoretical equations which define speed for the drive-bar type governor. Design considerations necessary for optimum speed regulation indicated by the theory were applied in establishing the shape and working relationship of various components in the drive-bar governor. Governors constructed to these dimensions have operated as forecast by the theory. The excellent agreement between theory and practice indicates that it is both desirable and practicable to apply the Moore theory in the design of governors.

The initial requirements imposed on the design of a governor for the 7-type dial were two-fold. The new governor had to provide speed regulation at least equal to the conventional fly-bar type and no objectionable noise could be created by the governor during operation. To better understand the reasons for noise in governors, suitable theory was developed for investigating this phenomenon. Application of this theory to any type governor results in an equation which defines chatter in terms of the constants of the governor. This equation and the equations determined by the Moore theory for speed regulation indicate the existence of an interrelationship between speed regulation and noise in governors. The theory indicates that noise free operation and good regulation are totally at variance. The fly-bar governor supports this conclusion since this governor having fine control of speed also produces a chattering noise during operation. To satisfy both requirements, good regulation and quiet operation, it is first necessary to design a governor which will regulate properly and secondly, if the constants selected indicate that chattering will occur, prohibit excessive noise by providing means for restraining the system during the unstable period.

This method of attack was taken in the design of the new drive-bar governor. By applying the Moore theory, a governor was developed for

use in telephone dials which is an improvement over the conventional fly-bar type governor. Fine regulation is provided by the drive-bar governor for a given change in the coefficient of friction between the studs and case. This is achieved by locating the studs as close to the weight pivot as manufacturing techniques will permit. Improved speed regulation is provided for varying input torque in the new governor as compared with the fly-bar governor. The experimental data shows the new design able to control speed approximately twice as well. It is an effective non-forcing governor, prohibiting excessive increase in dial speed as a result of forcing the fingerwheel during rundown. This nonforcing feature is achieved by applying the driving torque to the weights at a point to develop a moment about the weight pivot. This drive-bar moment assists the centrifugal force in maintaining pressure of the friction stud against the case for friction governing.

Having established a design which provided the degree of dial speed regulation considered necessary, it was then possible to investigate the second requirement of noise free operation. Application of the chatter theory to the drive-bar governor indicated the design to be unstable. This situation was controlled by using the ends of the drive-bar to limit the fall of the governor weights. This configuration of parts allows only small movement of the weights during the unstable period and provides damping as the weights close on the arms. The small motion and friction damping in the assembly results in a governor which is relatively free from noise during operation. Experience with dials equipped with drive-bar governors indicate that the chattering effect has been controlled.

ACKNOWLEDGMENT

The unpublished work of C. R. Moore covering the theory of fly-ball, fly-bar and band type governors has served as a foundation for this paper and the experimental work of R. E. Prescott aided considerably in determining the final drive-bar governor design. The writer also wishes to express appreciation to Mr. Prescott and H. F. Hopkins for their helpful comments and suggestions during preparation of this paper.

APPENDIX I

DERIVATION OF THE DRIVE-BAR TYPE GOVERNOR SPEED EQUATION

Referring to Fig. 4, as the governor mechanism rotates in a clockwise direction each weight tends to move outward under the influence of centrifugal force and the torque force, F . The centrifugal force, F_m ,

acts through the center of gravity of the weights radially from the turning center of the governor shaft. The torque force, F , is applied on the weights by the drive-bar arms. These forces are opposed by the spring force, F_s . A stud-to-case force, F_n , and a frictional component of this force, μF_n , act on the weights when the friction studs are in contact with the case. In deriving the equation of motion for speeds in excess of the critical velocity the following symbols will be used as noted on Fig. 4.

F — Force applied by torque on governor weights

F_n — Normal force of case acting on studs

F_s — Force exerted by spring when studs touch case

F_m — Centrifugal force acting at center of gravity of each weight

μ — Coefficient of friction

I_0 — Moment of inertia of the governor about center shaft

ω — Angular velocity of governor

ω_0 — Critical angular velocity at which studs just touch the case

m — Mass of each weight

r_0 — Radius to the center of gravity of each weight

r — Radius of governor case

α — Stud angle

Neg. Rotation — Rundown of governor

From the schematic, Fig. 4, taking moments about B we have

$$F_m b - F_s b + \mu F_n c - F_n d + F = 0 \quad (1)$$

collecting terms

$$b(F_m - F_s) - F_n(d - \mu c) + F = 0 \quad (2)$$

$$F_n = \frac{b(F_m - F_s) + F}{(d - \mu c)}$$

The driving torque on the governor is $G = 2Fec$; the retarding torque, $2\mu F_n r$. The difference between the driving torque and the retarding torque is as follows:

$$I_0 \dot{\omega} = G - 2\mu F_n r \quad (3)$$

where I_0 = Moment of inertia of the governor about the shaft center

$\dot{\omega}$ = Angular acceleration about shaft center

equating

$$F_n = \frac{G - I_0 \dot{\omega}}{2\mu r} \quad (4)$$

Combining equations (2) and (4) and solving for the acceleration, $\dot{\omega}$,

$$\frac{b(F_m - F_s) + F_j}{(d - \mu c)} = \frac{G - I_0 \dot{\omega}}{2\mu r}$$

$$\frac{2F_m b \mu r}{(d - \mu c)} - \frac{2(F_s b - F_j) \mu r}{(d - \mu c)} = G - I_0 \dot{\omega}$$

or

$$\dot{\omega} = \frac{G}{I_0} - \frac{2F_m b \mu r}{I_0(d - \mu c)} + \frac{2(F_s b - F_j) \mu r}{I_0(d - \mu c)}$$

Substituting the following values for F_m , F_s , F

$$F_m = m\omega^2 r_0$$

$$F_s = m\omega_0^2 r_0$$

$$F = \frac{G}{2e}$$

then

$$\dot{\omega} = \frac{G}{I_0} - \frac{2m\omega^2 r_0 b \mu r}{I_0(d - \mu c)} + \frac{2m\omega_0^2 r_0 b \mu r}{I_0(d - \mu c)} - \frac{j\mu r G}{I_0(d - \mu c)e}$$

Substituting for the design constants

$$K = r_0/r$$

$$M = 2mr^2 b K$$

$$\dot{\omega} = \frac{G}{I_0} - \frac{M\mu\omega^2}{I_0(d - \mu c)} + \frac{M\mu\omega_0^2}{I_0(d - \mu c)} - \frac{\mu r j G}{I_0(d - \mu c)e}$$

or

$$\dot{\omega} + \frac{M\mu}{I_0(d - \mu c)} \omega^2 = \frac{G}{I_0} + \frac{M\mu}{I_0(d - \mu c)} \omega_0^2 - \frac{\mu r j G}{I_0(d - \mu c)e} \quad (5)$$

This is of the form

$$\frac{d\omega}{dt} + g\omega^2 = h$$

or

$$dt = \frac{d\omega}{h - g\omega^2}$$

Separating variables and integrating*

$$t = \frac{q}{2h} \text{Ln} \frac{q + \omega}{q - \omega} + c \quad \text{where} \quad q^2 = \frac{h}{g} \quad (6)$$

Applying the initial conditions to solve for the constant c

$$\omega = \omega_0 \quad \text{at} \quad t = 0$$

$$c = - \frac{q}{2h} \text{Ln} \frac{q + \omega_0}{q - \omega_0}$$

Substituting in equation (6)

$$t = \frac{q}{2h} \text{Ln} \frac{q + \omega}{q - \omega} - \frac{q}{2h} \text{Ln} \frac{q + \omega_0}{q - \omega_0}$$

Letting

$$A = \frac{q + \omega_0}{q - \omega_0}$$

Then

$$t = \frac{q}{2h} \text{Ln} \frac{1}{A} \frac{(q + \omega)}{(q - \omega)} \quad (7)$$

and

$$\omega = q \tanh \left(\frac{ht}{q} + \text{Ln} \sqrt{A} \right) \quad (8)$$

Equation (8) applies as the equation of motion for the drive-bar governor for speeds in excess of the critical speed when

$$g = \frac{M\mu}{I_0(d - \mu c)} \quad (9)$$

$$h = \frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{I_0(d - \mu c)} \quad (10)$$

$$q = \sqrt{\frac{\bar{h}}{g}} = \sqrt{\frac{G(d - \mu c) + M\mu_0^2 - \mu r j G/e}{M\mu}} \quad (11)$$

APPENDIX II

GOVERNOR INPUT TORQUE

In order to apply the theoretical speed equations and the chatter equations developed for governors one must determine suitable values

* Short Table of Integrals, Pierce, B. O., pp. 8, No. 50.

for the stud-to-case coefficient of friction, μ , and governor input torque, G . Experimental evidence indicates a μ of 0.25 exists during normal operating conditions for hard rubber on brass. The values for governor input torque given in Tables I and III and used in the theoretical analysis for the governors were determined as follows.

The initial torque applied to the governor for the period up to the critical velocity was calculated from oscillograph string traces. These traces were obtained by mounting on the end of the governor shaft a thin disc having 36 radial slots spaced uniformly about the circumference. Light, detected through the slots of the rotating disc on the element of a photo tube, appeared as a distorted sine wave on the photographic paper. The distance between two successive wave peaks represented 10° of rotation of the governor. By noting on the trace the time between peaks, it was possible to determine the average velocity of the governor at 10° intervals after release of the fingerwheel, or start of rotation of the governor mechanism. The complete plot of these velocities appear as the experimental speed curve on Fig. 6. Inspection of the experimental curve for the drive-bar governor shows constant acceleration immediately after release. This appears as a straight line in the velocity time curve. Using the slope of this line and the moment of inertia, $I_0 = 7.4 \text{ gm cm}^2$, the initial governor torque was calculated as follows:

$$G_I = I\dot{\omega} = \frac{I_0}{t'} = \frac{100(7.4)}{55} = 13,480 \text{ dyne-cm.}$$

The governor torque value during normal rundown was found by first determining the governor stud-to-case force, F_n , and using this value in the equation

$$G = 2\mu F_n r$$

The fly-bar governor mechanism was used to determine the F_n force since the moment equation for this type governor contains measurable values. Referring to Fig. 7, the schematic of the fly-bar governor, the moment equation about B is as follows:

$$F_m b - F_s b - F_n d + \mu F_n c = 0$$

To solve this equation for F_n one must determine the centrifugal force, F_m , and the spring force, F_s . Using electronic flash equipment with an exposure time of $1/10,000$ of a second, it was possible to take distortion free photographs of the governor mechanism at the middle of the rundown. These photographs were taken with the governor in the horizontal

TABLE V — TYPE DIAL GOVERNOR INPUT TORQUE AT PULSE NO. 5

Dial No.	P.P.S.	F_m	F_s	F_n	G
1	9.89	37,820	32,820	15,320	9,050
	10.05	39,050	35,950	12,580	7,420
	9.92	38,100	34,300	11,650	6,880
2	10.02	38,920	35,230	11,340	6,700
	10.01	38,600	34,880	11,400	6,730
	9.96	38,400	35,000	10,420	6,160
3	10.03	38,950	34,100	14,850	8,770
	10.22	40,600	35,420	15,880	9,360
	10.08	39,400	35,000	13,522	7,980

Average 7,500 dyne-cm, approximately.

position, thus eliminating the gravity effect. The deflection of the governor spring measured on the photograph was used to determine the F_s force, and the governor speed, necessary to determine F_m , was taken from the oscillograph string trace. Three dials were tested, each having the same maximum motor spring torque, 490,000 dyne-cm, and clean governor cases and studs to produce an assumed coefficient of friction value, $u = 0.25$. For three 7-type dials with fly-bar governors, the experimental data given in Table V applies.

As determined, this 7,500 dyne-cm torque at the governor exists at the middle of rundown of the dial. In order to compare it with the 13,500 dyne-cm initial torque previously determined, it is first necessary to consider the effect of the motor spring. As stated previously, the torque provided by the motor spring during dial rundown decreases approximately 35 per cent from the initial value of 490,000 dyne-cm. Logically, the torque at the governor would decrease by the same percentage. Applying this factor to the torque value for the middle of rundown gives a value of 10,500 dyne-cm which can be compared with the 13,500 dyne-cm torque. A difference of 3,000 dyne-cm exists for the value of initial torque at the governor as determined by the two test methods. This remaining difference can be explained by considering the frictional losses in the dial mechanism during rundown. This analysis follows:

During rundown of the dial mechanism, a pair of pulsing springs tensioned against components on the pulsing shaft are alternately raised and lowered. This action allows contacts on the springs to open and close for pulsing in the telephone line. A portion of the input torque provided by the motor spring is required for performing this function. On Fig. 14 are plotted the torque curves for these pulsing springs as the pulsing

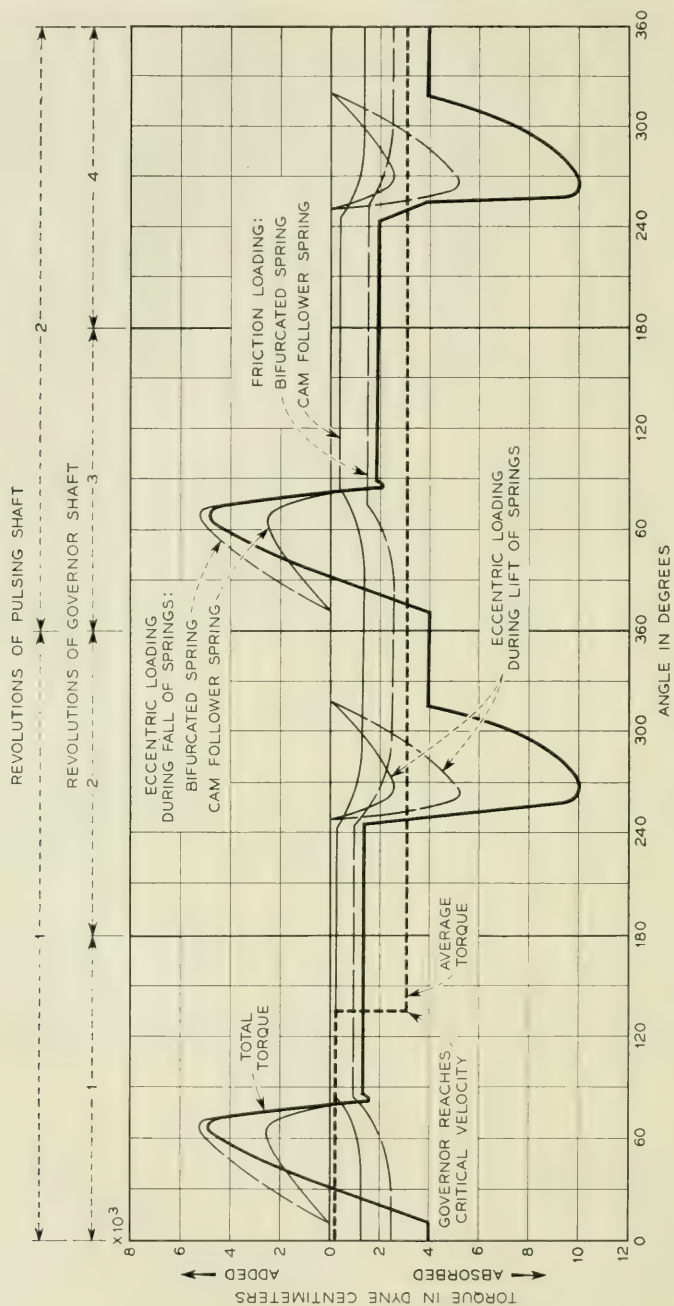


Fig. 14 — Torque analysis for the 7-type dial pulsing shaft.

shaft rotates during rundown. Fig. 15 is a schematic of the pulsing springs as they appear when the contacts are closed and when open.

For a short period during each revolution of the pulsing shaft, the pulsing springs aid the motor spring in driving the gear system. This occurs when the springs are being lowered by the cam just prior to opening of the contacts. For the remaining portion of each revolution the motor spring must provide energy to overcome frictional losses and lift the springs. These changes in energy required for moving the springs have been combined to give the total instantaneous torque curve also shown on Fig. 14.

As indicated by the pulsing spring torque analysis, the pulsing mechanism absorbs an average of only 200 dyne-cm during the period when accelerating up to the critical velocity. For rotation after the critical velocity, the average torque needed to drive the pulsing mechanism is approximately 3,000 dyne-cm. The difference between these average torque values appear at the governor shaft as a 1,500 dyne-cm torque difference. That is, 1,500 dyne-cm more torque is available for driving the governor prior to the time the critical velocity is reached as compared to that available after this time. This accounts for one half of the 3,000 dyne-cm difference in initial governor torques as calculated by the

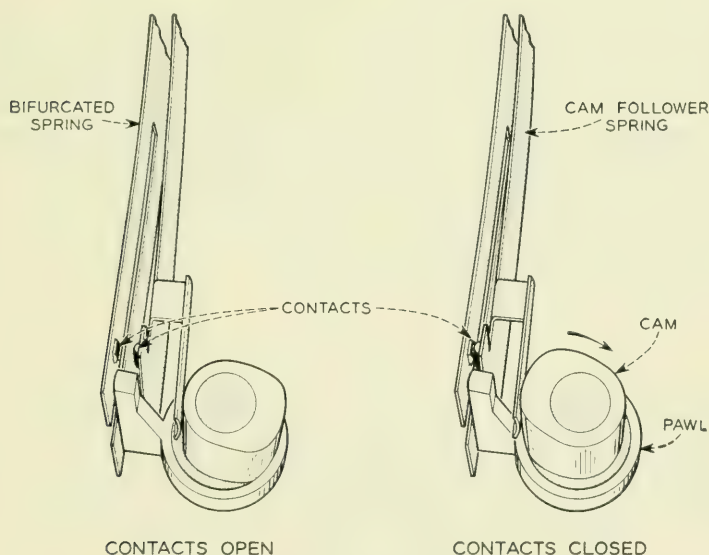


Fig. 15 — Pulsing springs of 7-type dial.

TABLE VI

Average torque at governor — Test No. 1, up to the critical velocity		13,500
Average torque at governor — Test No. 2, after reaching the critical velocity	10,500	
Torque at pulsing mechanism	1,500	
Torque required to overcome friction in the bearings	1,500	
Total	13,500 dyne-cm	13,500 dyne-cm

two test methods. The remaining 1,500 dyne-cm difference can be accounted for by considering the friction in the mechanism before and after the critical velocity.

Initially the system is accelerating as a simple fly wheel under the influence of the motor spring. At the critical velocity the governor studs engage the case and the dial rotates at virtually constant speed for the rest of the rundown period. This implies that the average speed of the moving parts in the mechanism will be twice as great for the period after the critical velocity as before. By considering the friction which exists in the dial bearings,* one can see the effect of this change in speed on the torque required to drive the mechanism. In sliding bearings with film lubrication the coefficient of friction is a function of speed. Specifically, as the speed of rotation of the journals increases, the coefficient of friction in the bearings will increase. Since the regulated dial speed is greater than the average speed while accelerating, friction in the system will also be greater. One can, therefore, justify the remaining 1,500 dyne-cm torque difference which exists before and after the critical velocity by considering it to result from the increased friction at the higher speed.

Therefore, by considering the effect of the pulsing mechanism and friction in the system, it has been possible to account for the difference in torque determined by the two test methods. Table VI shows the disposition of the torque.

APPENDIX III

DERIVATION OF CHATTER EQUATION FOR FLY-BAR TYPE GOVERNOR

Consider the governor rotating in the vertical plane at constant speed ω . Referring to the schematic Fig. 7 and taking moments about B

$$F_m b - F_s b - F_n d + \mu F_n c - m l \sin \beta = 0$$

* Design of Machine Members, Valence and Doughtie, p. 255.

where m = mass of weight

l = distance from C.G. of the weight to the pivot (B)

β = operating angle of governor weights with respect to a horizontal plane

For chattering to occur we know that the weight must leave the governor case. Therefore, at some angle β the gravity component will equal the forces tending to move the weight outward. For this condition pressure of the friction stud against the case will be equal to zero and

$$F_n d = 0$$

$$\mu F_n C = 0$$

and the moment equation becomes for this equilibrium condition

$$F_m b = F_s b + ml \sin \beta \quad (1)$$

Centrifugal Moment = Spring Moment + Gravity Moment

By applying the steady-state speed equation and the equation for centrifugal force we can transpose equation (1) to contain only design constants of the governor. From the steady-state speed equation

$$\omega = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2}{M\mu}}$$

Substituting $F_s = m\omega_0^2 r_0$ and $M = 2mrr_0b$

$$mrr_0b\mu\omega^2 = \frac{G}{2} (d - \mu c) + \frac{mrr_0b\mu F_s}{mr_0}$$

or

$$F_s = mr_0\omega^2 - \frac{G}{2} \frac{(d - \mu c)}{rb\mu}$$

also, the centrifugal force is

$$F_m = m\omega^2 r_0$$

Substituting in equation (1) the values for (F_s) and (F_m)

$$m\omega^2 r_0 b = mr_0\omega^2 b - \frac{G(d - \mu c)}{2r\mu} + ml \sin \beta$$

or

$$ml \sin \beta = \frac{G(d - \mu c)}{2r\mu}$$

where the gravity moment ($ml \sin \beta$) must be just equal to or smaller than $[G(d - \mu c)]/2r\mu$ to have no chatter occur in the governor. Expressing this in terms of the mass of the weight we have the chatter equation

for the fly-bar governor as

$$m \leq \frac{G(d - \mu c)}{2r\mu l \sin \beta} \quad (2)$$

APPENDIX IV

DERIVATION OF CHATTER EQUATION FOR DRIVE-BAR TYPE GOVERNOR

Referring to the schematic, Fig. 4, consider the governor rotating at constant speed ω in the vertical plane. Taking moments about B

$$F_m b - F_s b + F_j + \mu F_n c - F_n d - ml \sin \beta = 0$$

where m = mass of weight

l = distance from C.G. of weight to pivot B

β = operating angle of governor weights

Assume that at some angle β the gravity component will be large enough to make the F_n force equal to zero and a condition of equilibrium exists. For this condition

$$F_n d = 0$$

$$\mu F_n c = 0$$

and therefore, F_j , the torque component on the weight, must also equal zero. The moment equation becomes

$$F_m b = F_s b + ml \sin \beta \quad (1)$$

Using the equation for centrifugal force

$$F_m = m\omega^2 r_0$$

and steady-state speed for the drive-bar governor

$$\omega = \sqrt{\frac{G(d - \mu c) + M\mu\omega_0^2 - \mu r j G/e}{M\mu}}$$

where

$$F_s = m\omega_0^2 r_0 \text{ and } M = 2mrr_0b$$

Solving for (F_s)

$$F_s = mr_0\omega^2 - \frac{G}{2r\mu b} (d - \mu c - \mu r j/e)$$

Substituting in equation (1), the values for F_m and F_s

$$m\omega^2 r_0 b = mr_0\omega^2 b - \frac{G}{2r\mu} (d - \mu c - \mu r j/e) + ml \sin \beta$$

or

$$ml \sin \beta = \frac{G}{2r\mu} (d - \mu c - \mu r j/e)$$

Expressing the equation in terms of mass of the weight we have the chatter equation for the drive-bar governor.

$$m \leq \frac{G(d - \mu c - \mu r j/e)}{2r\mu l \sin \beta} \quad (2)$$

In-Band Single-Frequency Signaling

By A. WEAVER and N. A. NEWELL

(Manuscript received June 7, 1954)

Single-frequency signaling liberates dial systems from the restrictions of dc signaling methods. This freedom, as might be expected, is most important in the long distance telephone plant where trunks are frequently too long or have no conductors for dc signaling. The general plan of signal frequency (SF) signaling is based upon continuous signaling because of its speed and reliability. In this respect it is like the usual dc trunk signaling schemes. SF uses steady current in the trunk signaling path for the normal idle trunk condition and no current in the signaling path for the other and alternate busy (talking) trunk condition. This choice of signal conditions is essential for SF signaling in-band systems, which as the name implies operate within the standard voice channel, to avoid conflict between signal and voice transmission. The same conditions are also used in SF out-of-band and separate line systems.

The in-band SF system can be used with any type or length of line facility that meets normal voice transmission requirements and is therefore the preferred method used by the Bell System to meet requirements for toll dialing on a national basis, with other signaling arrangements limited to the shorter trunks. The requirements, design considerations, main features, and method of operation for the in-band system are outlined in this paper.

INTRODUCTION

The signaling requirements for dial telephone operation are naturally more exacting than those for manual switching methods. This means a high order of signaling system is needed to satisfy the requirements for the toll telephone plant and for automatic toll switching systems described in recent papers in this JOURNAL.^{1, 2} Indeed the advantages in speed and economy of dial telephone systems depend to a large extent upon the type of signaling provided for them. The signaling arrangements for intertoll telephone trunks which are the links between telephone switching systems, therefore, become most important.

Dial operation in the past has been based upon dc signaling which is

limited to relatively short distances and to line facilities having dc paths available to them. In planning for nationwide dialing of long-distance calls the need for an ac signaling system for dial telephone trunks became apparent. The length of intertoll trunks and the extensive growth in carrier line facilities which do not have associated dc paths made it necessary to develop ac dial signaling systems. The single-frequency signaling plan was developed for this purpose and is the first of its kind to satisfy the conditions associated with long distance intertoll dialing in the Bell System.

There are now several trunk signaling means that may be grouped as using the SF signaling plan. These are (1) the adaptation of VF carrier telegraph requiring an additional line channel independent of the voice transmission line facilities, (2) N1 and O1 carrier signaling furnished as part of these carrier terminals using 3,700 cycles outside but adjacent to the voice paths, and (3) 1600-cycle and 2400-cycle signaling systems, the in-band systems that use the voice paths.

Both in-band and out-of-band signaling have advantages and disadvantages. The in-band single-frequency signaling system uses ac in the voice frequency range to pass full supervision and dial pulsing signals over the same paths that are furnished for voice transmission in telephone trunks. This is accomplished without any loss in band width, change in line facility or addition of intermediate signaling equipment. On most calls voice and signal transmission are not required at the same time. On the few calls going to intercept operators voice transmission is impaired slightly by the effect of signal tone being on in one direction. On calls encountering busies, it is desirable to return both flashing supervision and interrupted audible tones. This can be done with out-of-band signaling but in-band signaling can return either but not both. The signaling system allows remote build-up and breakdown and provides for supervision of the temporary connections ordinarily used. Control and supervision of distant ends of trunks is required continuously whereas dial pulsing is required only at the start of calls and speech transmission is required only when connections are established.

TRUNK SIGNALS

Before going into the details of the signaling system itself it seems appropriate to review the trunk signals it is called upon to transmit. Most intertoll trunks are arranged for two-way operation, which means that a connection can originate at either end. To permit this operation, the signaling in each direction must be symmetrical and the trunk must allow the direction in which the connection is established to determine

TABLE I — CALLING TO CALLED DIRECTION

Trunk Condition	Signaling Frequency
Idle (disconnect)	On
Connect	Off
Dial pulsing*	On, then off, on pulses corresponding with dial break intervals
Ring forward	On, then off, one pulse
Disconnect (idle)	On

* Multifrequency pulsing,³ a separate a-c signaling system, is a faster means of transmitting number information often used instead of dial pulsing. Its use eliminates only the dial pulsing signals.

TABLE II — CALLED TO CALLING DIRECTION

Trunk Condition	Signaling Frequency
Idle (on-hook)	On
Stop pulsing*	Off
Start pulsing*	On
Flashing†	Off, then on, off pulses corresponding with off hook supervision
Off-hook (answer)	Off
Ring back	On, then off, on for duration of ring
On-hook (idle)	On

* Stop- and start-pulsing control signals are required only in connection with common control switching equipment.

† Flashing supervision signals are required only for operators; the on intervals light cord circuit lamps to inform operators of the status of calls independently of the position of cord circuit talking keys.

the signaling interpretation. The latter is conveniently identified by different names for the trunk signals in the two directions. Only two signal conditions, that is, tone on or tone off, in each direction are required for all dial trunk signals. Continuous dependence upon these two conditions assures a high degree of reliability because of signal redundancy. Tables I and II show the required dial intertoll trunk signals, together with the action taken in regard to the signaling frequency.

The signaling system must be able to handle minimum and maximum length signals. The minimum times occur in dial-pulsing where the shortest signal element may be as low as 30 milliseconds. All other types of signals have longer durations.

The maximum permissible transmission time for signals between trunk terminals is determined by the allowable unguarded intervals on two-way trunks, during which double connections may occur, and also by the stop-pulsing signal recognition interval. This time is limited to about 175 milliseconds.

The distortion permitted in the transmission of signals is proportional

to the time duration of each signal. In general, variations in signal time should be within ± 5 per cent. All effects of the trunk signal medium should be confined within the trunk terminals or be of such character as to have no adverse reaction in connected circuits. This is necessary for proper operation of switched connections.

BASIC PLAN

The in-band single-frequency signaling system, although fairly complex in detail, is very simple in principle. Normally, i.e., when the circuit is idle, steady tone is transmitted over the line and holds relays operated at the receiving end. Signaling is accomplished by removing and reapplying this tone, which in turn releases and reoperates the distant relays. Independent operation is obtained in each direction with one signal frequency on four-wire lines, which have separate one-way transmission paths from terminal to terminal, and with two signal frequencies, one for each direction of transmission, on two-wire lines.

The signaling system is provided as a separate entity. It is connected in series with the transmitting and receiving branches of the line circuit at each end of a trunk and to the terminal relay circuit (trunk circuit) by two one-way signaling leads. A typical arrangement for a four-wire line terminating in a two-wire switching office at the West terminal and a four-wire switching office at the East terminal is shown in Fig. 1.

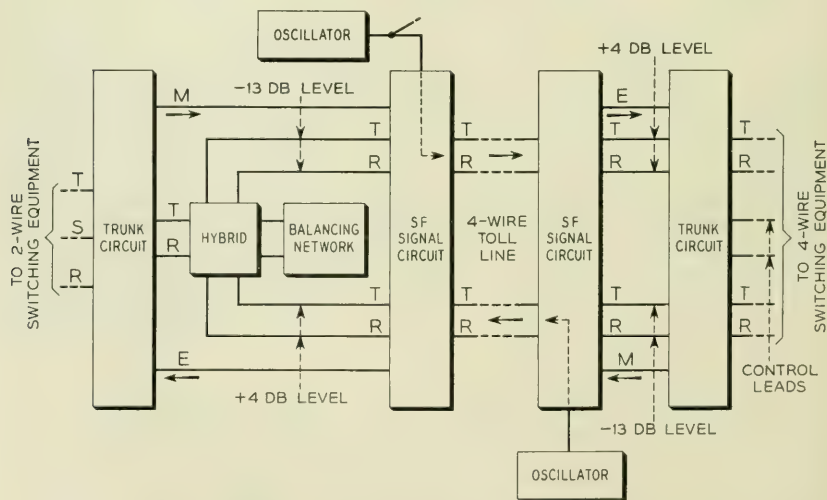


Fig. 1 — Application of single-frequency signaling to trunks with four-wire lines.

In the case of two-wire lines, the signaling equipment is applied to the four-wire transmission paths of terminal repeaters, using a different frequency for signaling in opposite directions. Band elimination networks are provided ahead of each receiving circuit to block the transmitting frequency, which would otherwise come into the receiver via echo paths and interfere with its operation.

GENERAL DESIGN CONSIDERATIONS

The successful use of the voice path for signaling, especially for continuous as contrasted to "spurt" signaling, is feasible only by a compromise among a number of conflicting factors. These factors or design considerations are (a) choice of signal frequency, (b) signaling power and receiver sensitivity, (c) imitation of signal by speech or tones, (d) interference to signal by other tones and noises, and (e) audibility of signaling tone to operators and subscribers.

(a) Choice of Signal Frequency

The choice of signal frequency is determined mainly by considerations of signal imitation by speech. As will be shown later on, signal imitation decreases rapidly as the signaling frequency is raised with the result that the highest frequency that can be reliably handled by the transmission path is used. In the case of some four-wire type lines, such as EB carrier, the highest frequency that should be used is 1,600 cycles. However, the use of 1,600 cycles results in an expensive signaling system and it is desirable to have another system using a higher frequency (2,600 cycles) for application to lines that can handle this frequency. These systems are basically the same in principle and both are described in the present article.

For application to two-wire lines the second frequency used is 2,000 cycles in the case of the older 1,600-cycle system and 2,400 cycles in the new 2,600-cycle system.

(b) Signaling Power and Receiver Sensitivity

To limit cross talk into adjacent voice channels and to avoid adding much signal power to the repeaters it is desirable to use the lowest practicable signal power consistent with a usable signal-to-noise ratio. A value of -20 dbm referred to zero transmission level for the steady idle tone is satisfactory for this purpose. In order to obtain an overall margin of 8 db the sensitivity of the receiver is set at -28 dbm. A higher power

is used for a short time at the beginning of each application of signaling tone to help overcome line noise and attenuation variation. This power increase is 14 db for the 1600-cycle system and 12 db for the 2,600-cycle system.

(c) Imitation of Signal by Speech or Plant Tones

An in-band signaling system requires that the receiver respond to signaling tone and at the same time be non responsive to speech formed currents. The principal design factors employed to achieve this feature are (1) the use of "guard action," (2) the employment of as narrow a bandwidth as practicable for the signal selective network, (3) the use of volume limiting, (4) the use of the longest operate time practicable, consistent with trunk signaling requirements, and (5) the use of the highest frequency that can be handled by the voice path.

"Guard action" is the principal means used in protecting the receiver against operation on speech. It consists in the use of nearly all frequencies in the voice band other than those in a narrow band centered on the signaling frequency to generate a voltage which is used to oppose that resulting from the signal frequency. The sum of these two voltages, plotted against frequency, for a typical receiver is shown in Fig. 2. A term used to specify the magnitude of the guard action is "guard-signal ratio" (G/S in Fig. 2) or just "guard ratio." The amount of guard which can be used is limited by signal to noise ratio because noise, like speech, tends to oppose operation of the receiver. A guard ratio in the range of 6 to 10 db has been found to be practicable.

Protection against signal imitation is also provided by narrowing the signal frequency band as much as practicable, since this reduces the effective operating power of voice and noise frequencies. However, the extent of this narrowing is limited since the operating bandwidth must be sufficient to allow for frequency variation in the signal supply, for carrier shift in the transmission path, for variation in the elements of the tuned circuit in the receiver and to allow for the transmission of the needed side bands of the signaling pulses.

A bandwidth of 60 to 75 cycles at the 3 db points at dialing power (about — 6 dbm at zero level) has been adopted as about the minimum that is practicable. Because of limiting and guard action the effective bandwidth is a function of input power and in the particular designs adopted approaches about 150 cycles at the just operate point.

Volume limiting is another means used to help prevent false operation on high levels of speech. The explanation of this action is illustrated in

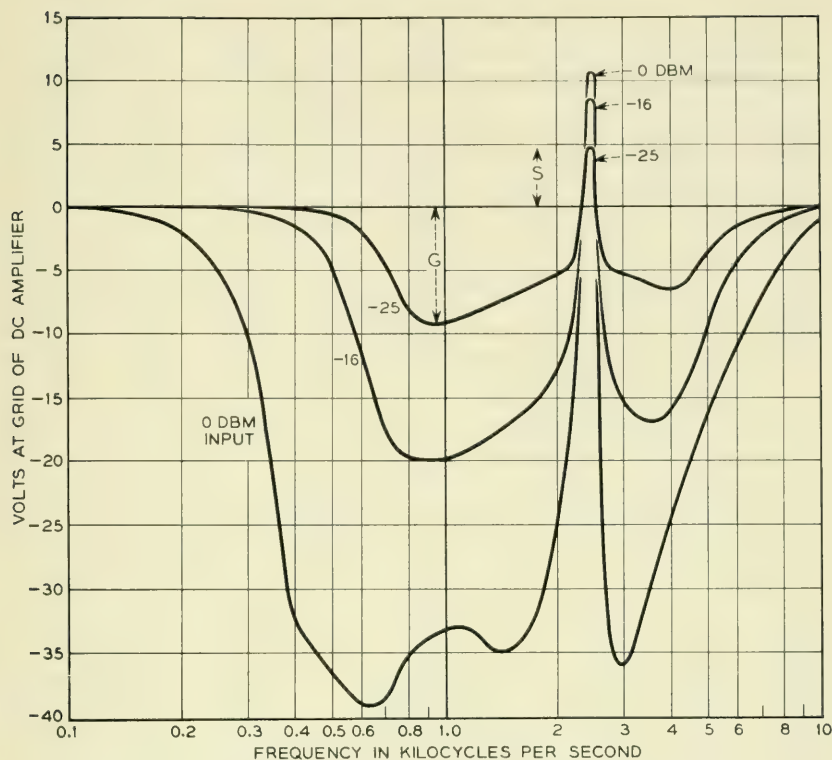


Fig. 2 — Signal-guard characteristics of 2,600-cycle receiver.

Fig. 3. The dotted line shows a characteristic for a receiver with no limiting, while the solid lines are for one with limiting. As shown, a given large input would produce an output of E_2 , the difference between the signal voltage and guard voltage components, for the former case and an output of E_3 , which is about half as much, for the latter case. This will be less likely to operate the receiver when applied for a short interval of time, although either will produce an operation if applied long enough, because either exceeds the just operate value E_1 .

Having established the basic design parameters of sensitivity, bandwidth, guard to signal ratio, limiting characteristics and speed of response it is important to know the relationship between signal imitation and the frequency used for signaling.

To obtain information on this subject a series of tests were made using a number of guard channel receivers as nearly alike as possible except for the frequencies of signal response, which were 800, 1,350, 1,800, 2,400

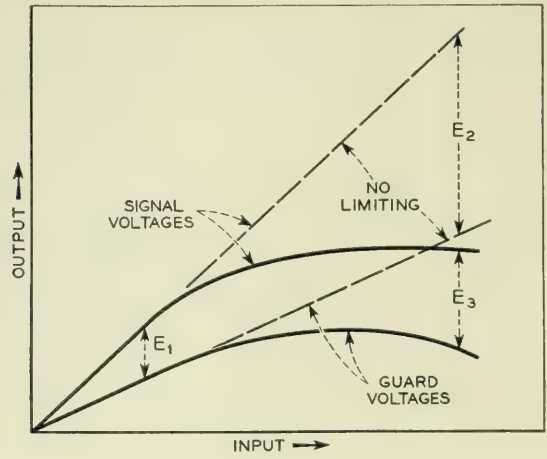


Fig. 3 — Limiter characteristics.

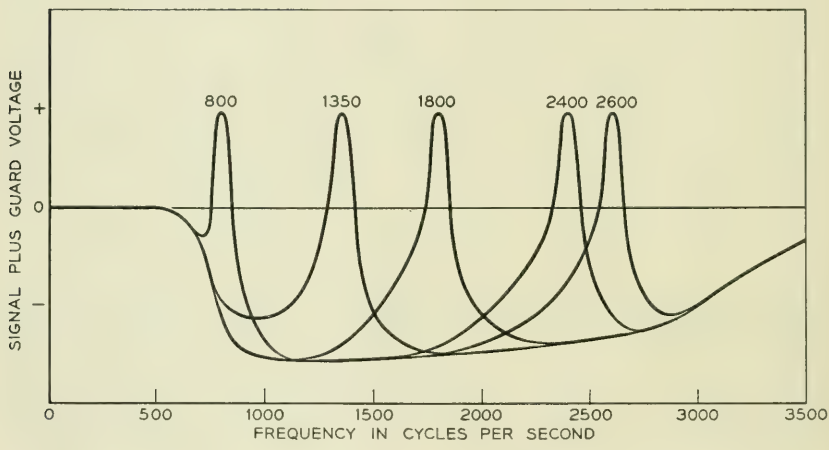


Fig. 4 — Signal-guard characteristics used in signal imitation tests.

TABLE III

Receiver sensitivity (0 level).....	-28 dbm
Receiver signal bandwidth (3 db points), at just operate level " approx.....	150 cycles
Guard to signal ratio.....	6 db
Start of limiting above just operate, approx.....	5 db
Minimum duration signal for operate.....	50 ms

and 2,600 cycles. The signal-guard characteristics of all of these receivers are shown in Fig. 4. Other parameters used are given in Table III.

The results of these tests are shown in Fig. 5, where frequency is plotted against signal imitations per 100 calls. The receivers were located at New York and were connected at different times in trunks to Boston, Toronto, Buffalo, Pittsburgh, Washington and Miami, so as to get some geographical speech distribution. There was no detectable geographic effect.

The type of speech sound causing the signal imitation is also of interest even though we have not as yet been able to put this information

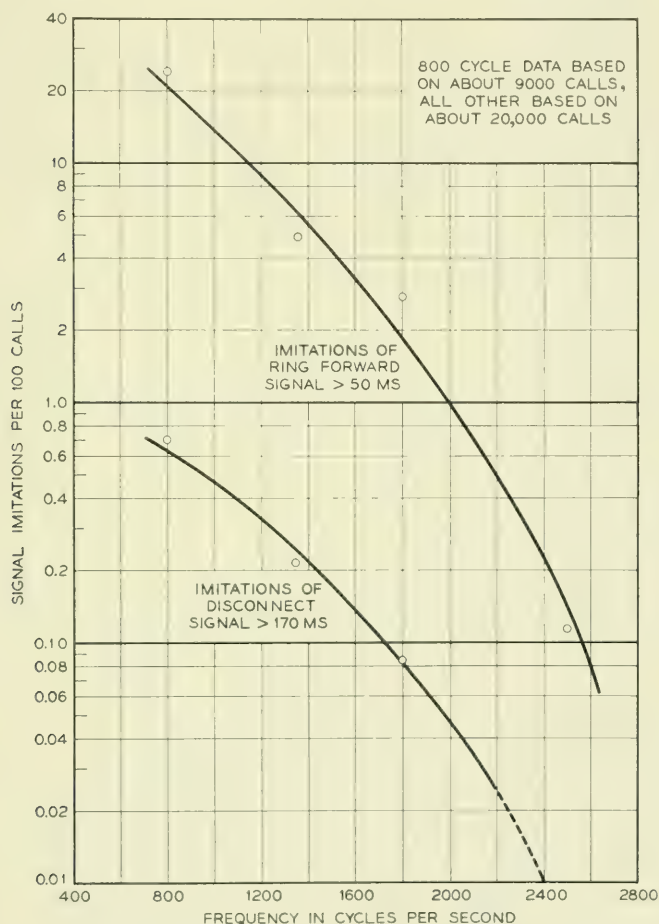


Fig. 5 — Results of signal imitation tests.

TABLE IV

Circuit "singing" or momentary circuit transient.....	10
Adult whistle.....	2
Uncertain.....	8
Tone spurt.....	1
Speech imitations.....	48
Total.....	69

to use in the design of single frequency guard type receivers. In observations on many thousands of calls it was noted that vowel type sounds were the predominant cause of signal imitations, with all except a few being formed by female speech. At the highest frequency tested (2,600 cycles) over 90 per cent were caused by the long *e* vowel sound (as in *feet*). At the intermediate frequencies 1,350 and 1,800 cycle) most vowel sounds were noted, while at 800 cycles signal imitations were caused principally by two sounds, namely *O* (as in *hole*) which accounted for about 50 per cent of the total and *aw* (as in *awl*) and similar sounds like *ah* as in *father*.

Signal imitations from vowel sounds are to be expected because of their relatively large energy and sustained nature. For instance it is well known⁴ that a sustained long *e* sound can have a large component in the high frequency range with very little energy in the range from 500 to 2,000 cycles where the guard action is effective. Likewise a sustained long *O* sound can have a large peak in the 500-cycle to 1,000-cycle range with little energy in the 1,000-cycle to 3,000-cycle range where the guard action is effective for the 800-cycle receiver.

Speech formed currents are not the only source of signal imitation. In one series of observations using 2,600-cycle receivers involving circuits from New York to a number of other cities including Toronto, Boston, Baltimore, Washington and Miami a total of 69 signal imitations were observed. In each case an attempt was made to determine the sound that caused the false operation, with the results given in Table IV.

NOISE CONSIDERATIONS

Noise affects the signaling system in a variety of ways depending upon the nature of the noise and upon the particular signaling function being performed. When tone is first applied it is of course desired that the receiver operate. However at this time the "guard" circuit is functioning because it is also desired that the receiver be non responsive to speech. Noise, which acts on the guard circuit like speech, will therefore tend to prevent operation of the receiver. If the noise is steady and large enough

it would of course permanently prevent operation, while if it is of short duration and occurred at the beginning of a signal interval it would only delay operation. Even a short delay would be harmful to ring forward or dialing signals but could be tolerated in disconnect or flashing signals.

An example of how a short duration high level noise affects receiver operation is illustrated in Fig. 6. As can be seen this particular noise transient, which happened to be caused by a relay in the trunk circuit, occurred just prior to and overlapping the tone interval. It would seriously damage a ring forward signal. The solution to this particular problem is to absorb the noise at its source, or prevent it from reaching the voice path.

After the receiver is operated for a short time (0.2 sec or so) the guard action is removed. Later on when the tone is removed it is desired that the receiver release, but noise at this time will tend to prevent release. The solution to this problem is a compromise in receiver sensitivity, i.e., it must be sensitive enough to hold up on the weakest signaling tone and yet release on the maximum noise that can be tolerated from a speech point of view. Fortunately such a compromise is achieved with a sensitivity that will cause the receiver to hold with the tone about 8 db

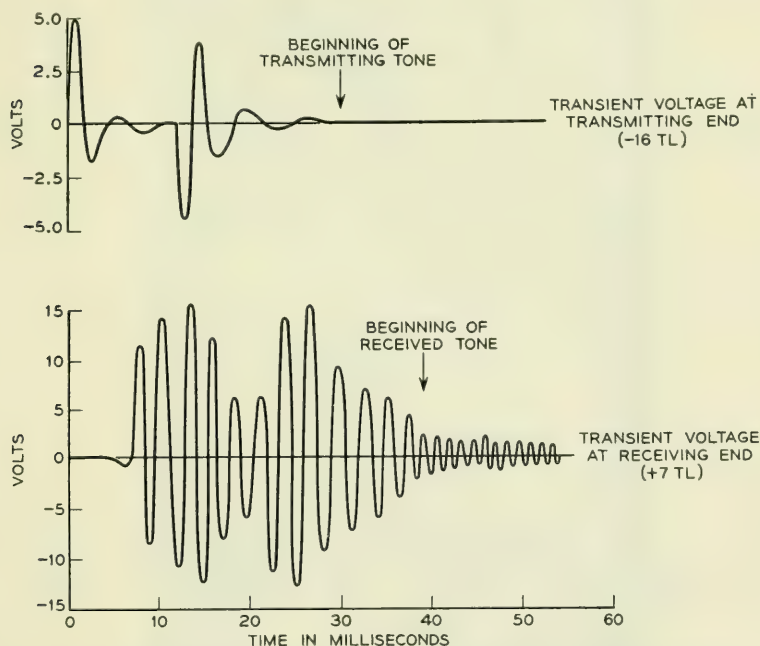


Fig. 6 — Example of transient voltages generated by relay operation.

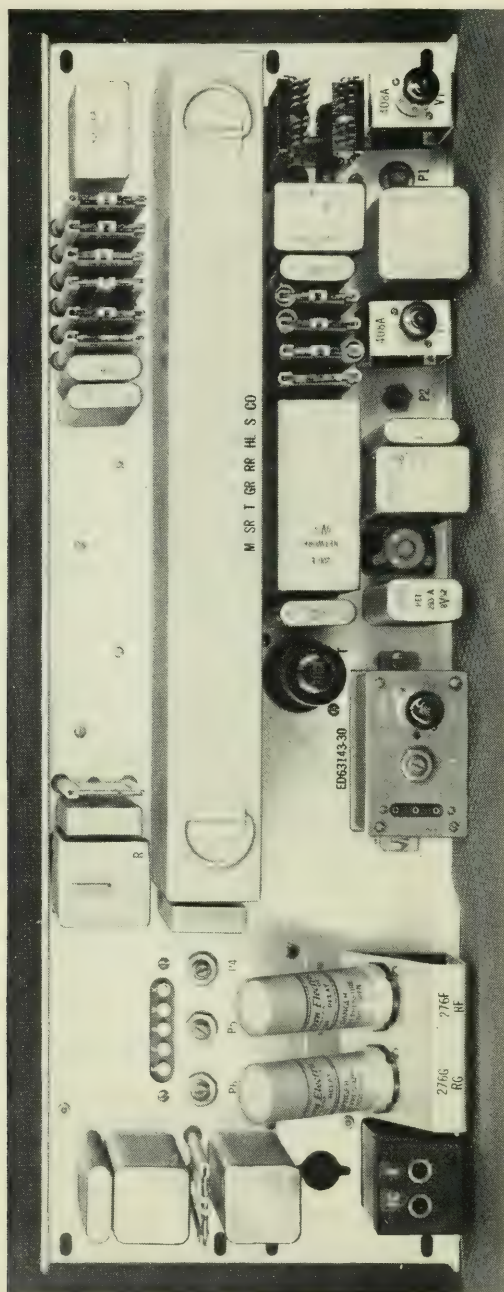


Fig. 7 — Front view of the 1,600-cycle signaling panel.

below normal and yet permit release in the presence of as much as 50 dba of thermal noise at zero transmission level.

SPECIFIC DESIGNS

The two signaling systems 1,600 and 2,600 cycles, are basically similar in principle. However, as can be seen by reference to Fig. 5, a simple guard type receiver having a frequency of 1,600 cycles would have too many signal imitations. To overcome this the guard ratio during the talking condition was increased from 6 to 10 db, the minimum signal interval to just cause a response was increased to 100 milliseconds during the talking condition and the sensitivity was decreased to -16 dbm. As a result fairly complicated timing and switching circuits are needed to assure that both transmitting and receiving circuits have the right condition at the right time.

Table V gives a summary of the principal design parameters of the two systems.

DESCRIPTION OF 1,600-CYCLE DESIGN

A front view of the 1,600-cycle main unit is shown in Fig. 7. This panel is 8 inches high by 23 inches wide and weighs about 20 lb. The essential elements of the circuit are shown in Fig. 8. It connects to the trunk relays over two leads E and M and into the line circuit via leads labeled T, R, T₁ and R₁. The transmitter, shown in the upper portion of the figure, uses dc biased germanium varistors (diodes) to control the application of signal current to the line, and for control functions, uses four relays designated M, CO, HL and RR. The functioning of the first three except for tone control are described under the heading "Description of 2,600 Cycle Design" later in this article. The RR relay (not shown) in conjunction with the M relay lengthens the sent pulse for the ring for-

TABLE V

	Dialing Condition		Talking Condition	
	1,600	2,600	1,600	2,600
Sensitivity, dbm.....	-28	-28	-16	-28
Bandwidth, cycles				
Low level.....	150	150	150	150
High level.....	60	75	60	75
Guard ratio, db.....	0	-6	10	6
Minimum signal for just operate, ms.....	35	35	100	50
Start of limit above just operate, db.....	5	5	5	5

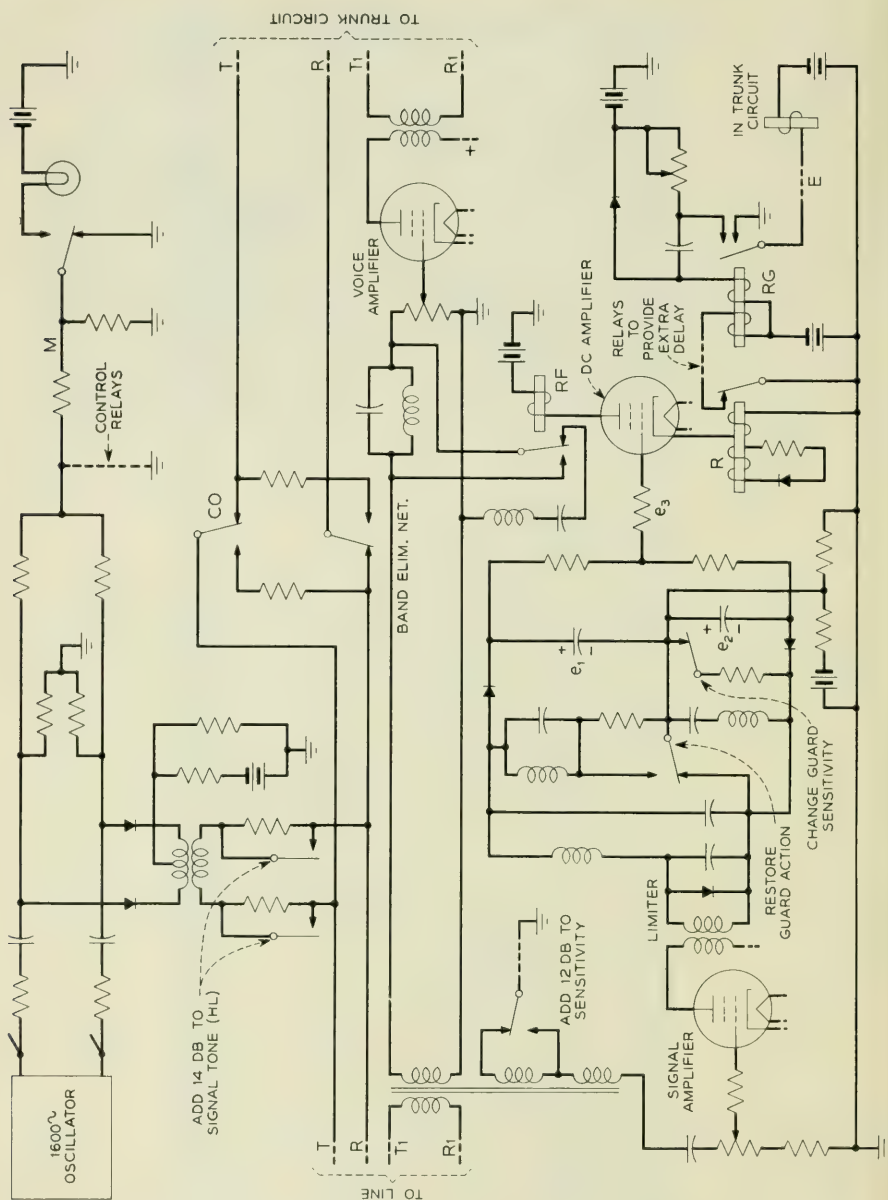


Fig. 8 — Simplified diagram of 1,600-cycle signaling circuit.

ward signal because the far end receiver at this time has a long operate time.

The signal receiver is connected in series with the receiving branch of the voice transmission path and is provided with a voice amplifier to provide a blocking function so noises originating in the switching equipment or beyond will not interfere with operation of the signaling receiver, and to compensate for the signaling bridging loss.

The receiving portion of the circuit is shown in the central and lower portions of Fig. 8. The idle condition of the trunk is shown, tone is being received, the R, RG and RF relays are operated and the band elimination filter is inserted in the receiving branch to prevent signaling tone from entering a connected circuit and interfering with signaling there.

The signal currents coming in from the line are passed through the signal amplifier, limiter and low pass filter and applied to the signal-guard network from which signal voltage is applied to the dc amplifier tube to operate the above mentioned relays and open the E lead, which extends into the trunk circuit. Typical wave forms at several points in the circuit are shown in Fig. 9. The extra operate time provided during the talking condition is obtained from slow relays (not shown) which at this time are in the path from the R to the RG relay. These relays also change the sensitivity, and guard ratio.

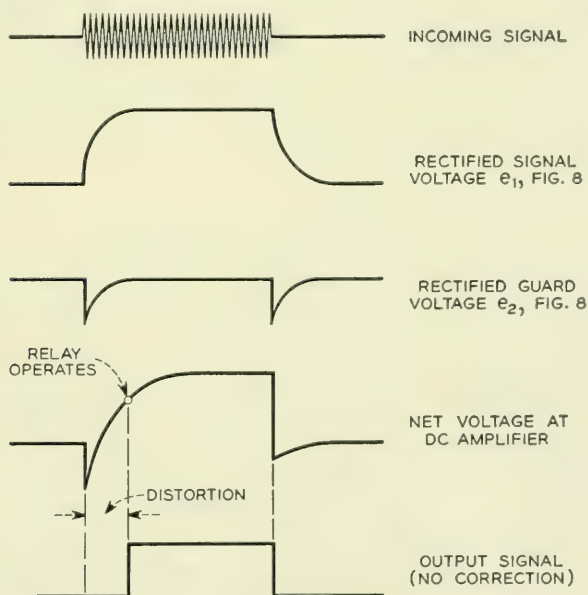


Fig. 9 — Typical wave forms in 1,600-cycle receiver.

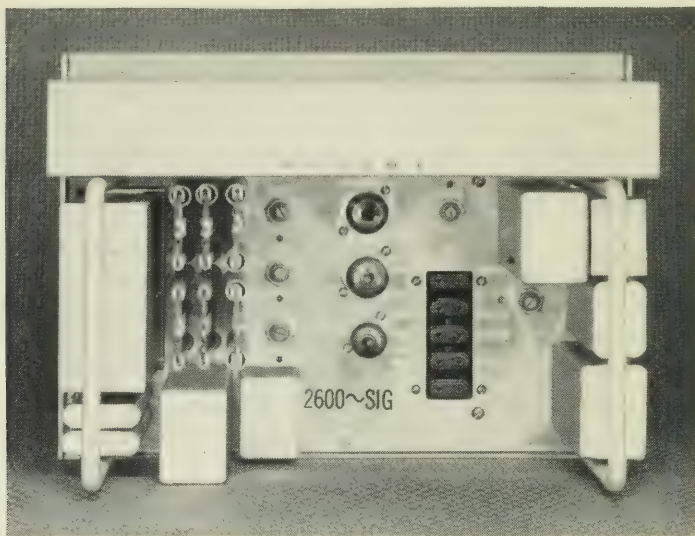


Fig. 10 — Front view of the 2,600-cycle signaling panel.

The R delay, because of guard action and the fact that its secondary winding is closed through a varistor and resistance, is relatively slow to operate and fast to release. For this reason some form of pulse correction is necessary to get good dial operation. This is obtained with the RG (regenerate) relay and its associated CR timing network, and there results an output pulse within the needed limits, even though the signal on the R relay is shortened considerably.

DESCRIPTION OF 2,600-CYCLE DESIGN

The 2,600-cycle unit, shown in Fig. 10, is just half the size of the 1,600-cycle unit, costs less than half as much, and is of the "plug in" type so it can be readily replaced for maintenance action.

A simplified diagram of the new signaling circuit is shown in Fig. 11, with the transmitting portion in the upper part of the figure and the receiver in the lower part. The transmitting portion employs three relays designated M, HL, and CO which are interconnected to perform the following functions: relay M is used to key the signaling tone; relay HL (high level) adds 12 decibels to the tone power at the beginning of each signal tone application to improve signal reliability in the presence of line noise and variations in attenuation; and relay CO (cut off) cuts the line momentarily to prevent noises originating in the switching equipment from interfering with signaling.

guard channel. These voltages are then applied to rectifying circuits where positive and negative dc voltages are developed and passed on to the dc amplifier tubes A and B, respectively. The R relay operates first and cuts in the band elimination filter thereby preventing signaling tones from entering a connected toll line and interfering with the signaling there. However, a short spurt of tone will get through because of the finite time required to operate this relay, and the R relay must therefore be made slow enough so that it will not operate on this tone. This action is obtained with the resistor-capacitor network (OT and C_3) in the grid circuit of the associated dc amplifier.

A short time (about 200 ms) after the R relay operates, a relay (not shown) releases, which short circuits the guard network and inserts enough resistance in series with the signal network to substantially remove its tuning. The purpose of removing the guard action during the idle condition is to prevent release of the receiver which would otherwise be caused by occasional bursts of line noise.

The signal network is made broad at this time for the following reason. In connections to an intercepting operator, "off hook" supervision is not returned to the originating end to avoid charging for the call. This means that tone remains on the line to continue to hold up the receiver. At the same time the intercepting operator's speech must of course be transmitted over the line so that both speech and tone enter the receiver.

Speech can be of a relatively high power as compared to the tone with the result that the action of the limiter tends to suppress the tone and could falsely release the receiver if the signal tuning were present. However, with broad tuning either speech or tone will hold up the receiver and no trouble is encountered.

The blocking amplifier seen in Fig. 11 has the same function as in the 1,600-cycle design described previously.

Among the new features in this unit, one of the most significant is the pulse-correcting circuit. This feature is a very important element in the entire long distance connection since it serves to keep the length of the dial pulses within specified time limits. The dial pulses on many calls may have to go through a number of central offices and all their associated equipment, and in each stage of the transmission path the ideal 60-millisecond dial pulse may be distorted so that it becomes too long or too short.

The pulse-correction is accomplished by generating appropriate transient voltages, whose duration is determined by capacitor-resistor networks. These voltages are then applied to the grid of dc amplifier B in Fig. 11 to perform the elongation or shortening of the pulses as required.

Dial pulses of 2,600-cycle tone with various degrees of distortion enter the circuit at the left of the diagram. It is desired that they be corrected, converted to an interrupted dc current in lead E by means of relay R, and then passed on through the toll office. This break in the dc current will effectively key the next outgoing tone transmitter, which faithfully reproduces the break intervals. The R and RF relays are shown operated as would be the case if an originally short dial tone pulse were being converted into a longer dc current break in lead E, i.e., closer to 60 milliseconds in duration.

Currents of signal frequency build up voltages across the "signal network," while currents of any other frequency, such as those from speech, build up voltages across the "guard network." These voltages are rectified separately, as shown, to obtain oppositely poled dc voltages across the signal capacitor S and the guard capacitor G. During normal speech, the voltages across the guard capacitor will dominate and, being negative, will keep the grid of dc amplifier B negative to prevent operation of the R relay at this time.

However, when dial tone pulses are applied, the main voltage will appear across the signal capacitor, with a small transient voltage appearing across the guard capacitor due to sidebands of the dialing frequency. This transient, being negative, produces an effect shown at (a)

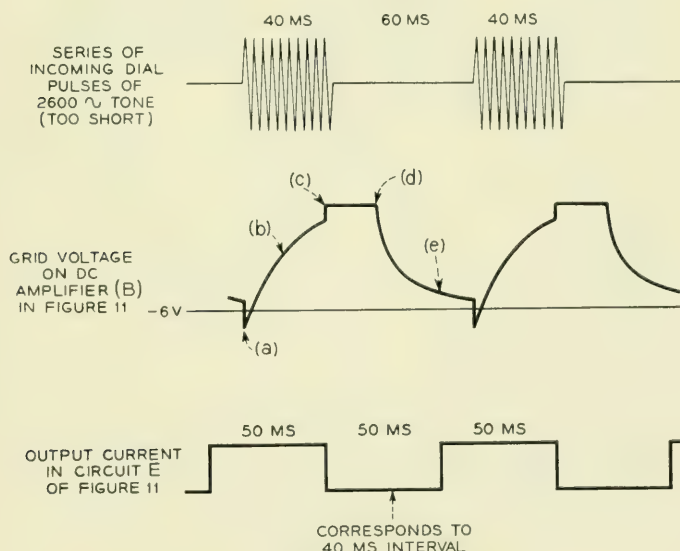


Fig. 12 — Wave forms for lengthening of short dial pulses.

in Fig. 12, the middle diagram of which represents the voltage on the grid of the dc amplifier B. The voltage on the grid of this amplifier will thereafter build up slowly as shown at (b) because of the or (operate time) resistor and c_3 capacitor. In this illustration it is assumed that the incoming dial pulse has a duration of only 40 milliseconds, which is shorter than desired. In this particular case the RF relay operates about at time (b) and the R relay at point (c), which is just prior to the end of the pulses. It is therefore necessary, in order to obtain a corrected output signal, to delay the release of the R relay. This is accomplished by causing current to flow into the grid of the dc amplifier tube B from time (c) to time (d). This current comes from energy stored in c_2 from battery D prior to time (c) and from the positive transient at the plate of tube A generated by the end of the pulse. At time (d) the RF relay releases, and the voltage on the grid of the dc amplifier B decays along the line from (d) to (e), at which point relay R releases. The result is that the R relay has sent on a pulse that has been corrected from an original 40 to a final 50 milliseconds. It is possible but uneconomical to build a unit that would achieve perfect pulse correction to 60 milliseconds, but when a number of toll lines are in tandem, the pulses will have to pass through

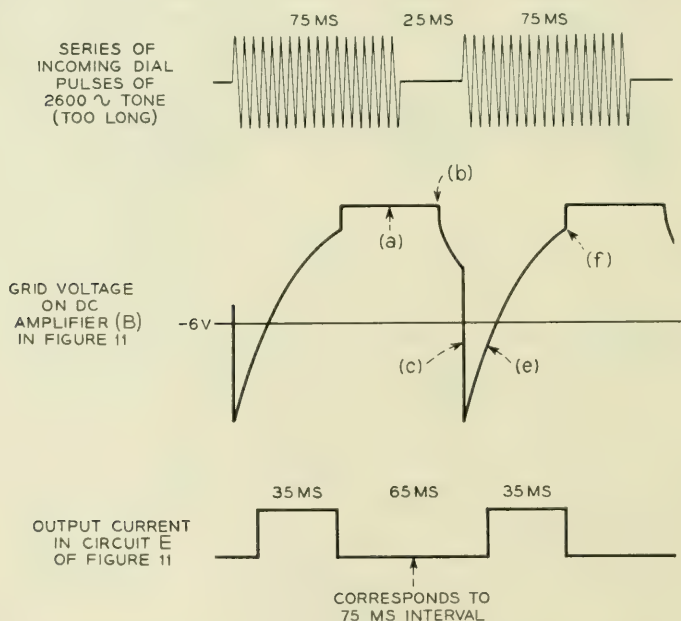


Fig. 13 — Wave forms for shortening of long dial pulses.

several SF units, and the pulses will therefore be brought close to 60 milliseconds by the successive corrections.

The action of the pulse corrector when the incoming signal is too long will be explained with the help of Fig. 13. Here the cycle of events will be assumed to start at (a), just at the end of an incoming dial pulse. From (a) to (b) current will flow into the grid of the dc amplifier B, at (b) the RF relay will release, and the voltage will start to decay, all actions so far being the same as in Fig. 12. In this instance, however, a new pulse would come along before the R relay has had a chance to release, and unless something is done about it, it would not release at all.

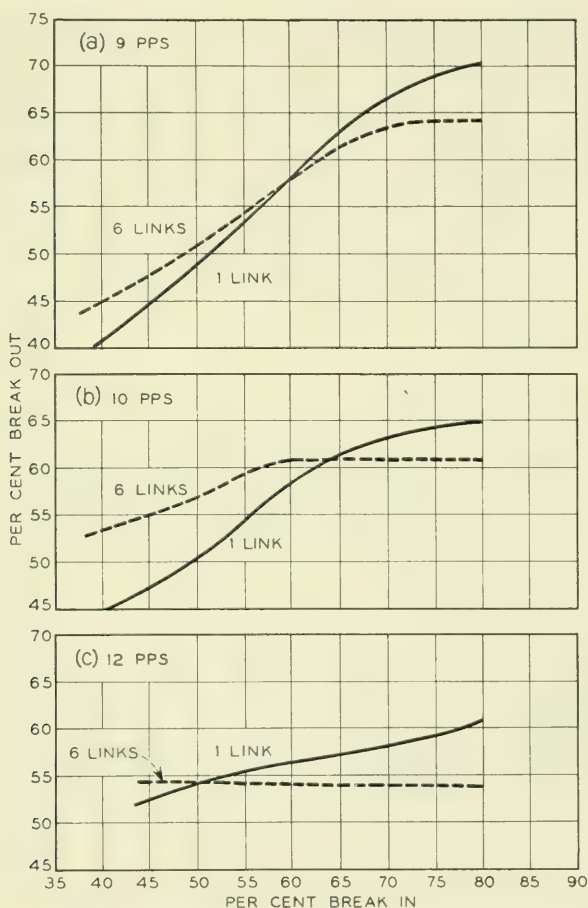


Fig. 14 — Per cent break input versus output characteristics for (a) 9 pulses per second, (b) 10 pulses per second, and (c) 12 pulses per second.

This "something" consists of using the large negative transient voltage generated at the plate of tube A resulting from the application of a positive pulse to its grid. This transient, shown at (c) drives the grid of B rapidly and heavily negative, thereby forcing the R relay to release. The remainder of the pulse-correcting action consists in using this same transient to delay the reoperation of the R relay. This is accomplished by storing some of the energy in the $R_1 C_1$ network. This slows the building up of the voltage as shown at (e), and the relay R reoperates at (f). The resultant repeated signal is shown below, where a 75-millisecond signal has been pulse-corrected to 65 milliseconds, further corrections being effected in the subsequent SF units.

Dialing performance of typical signaling units is shown by graphs in Figs. 14(a), (b), and (c). These curves show per cent break input plotted against per cent break output for 9, 10 and 12 pulses per second for one and 6-link operation. If the system were linear the input-output characteristic would be a 45 degree straight line. When the slope is less than 45 degrees there is pulse correction, and if the slope were zero with an output at 60 per cent, pulse correction would be perfect. It is noted that the pulse correction action improves as the speed increases and at 12 pulses per second the output is nearly independent of input.

ACKNOWLEDGEMENTS

The success of this project is the result of contributions by many people, and all cannot be named specifically. However special mention should be made of F. A. Hubbard, who designed the equipment arrangements of the 2,600-cycle system and W. W. Fritschi, C. W. Lucek, R. O. Soffel and A. K. Schenck who made important contributions to the circuit design.

REFERENCES

1. J. J. Pilliod, Fundamental Plans for Toll Telephone Plant, B.S.T.J., **31**, pp. 832-850, Sept., 1952.
2. F. F. Shipley, Automatic Toll Switching Systems, B.S.T.J., **31**, pp. 860-882, Sept., 1952.
3. C. A. Dahlbom, A. W. Horton, Jr., and D. L. Moody, Application of Multi-frequency Pulsing in Switching, A.I.E.E. Transactions, **68**, pp. 392-396, 1949.
4. H. Fletcher, *Speech and Hearing*, Van Nostrand.

Centralized Automatic Message Accounting System

By G. V. KING

(Manuscript received May 6, 1954)

A centralized automatic message accounting system (CAMA) has been developed so that the billing data can be recorded at a centralized crossbar tandem office for message unit and toll calls originated by telephone customers served by a large number of local dial central offices. It is an essential part of facilities for economical nationwide customer dialing through central offices with older types of switching equipment and through other central offices which could not otherwise economically give this service. The new system records the billing data on paper tapes in the same form now used by local automatic message accounting systems. Tapes for both local and centralized automatic message accounting systems are processed in the same accounting center.

INTRODUCTION

One broad objective of the Bell System is to extend the customer's dialing range so that ultimately he will be able to dial his own calls to any telephone in the country in much the same way as he now dials local calls. Several steps toward this goal have already been taken. A revised fundamental plan¹ for automatic toll switching² has been adopted which involves among other things the use of a nationwide numbering plan³ covering the United States and Canada. In accordance with this plan each customer will be given a distinctive 10-digit designation which will consist of a 3-digit regional or area code, a 3-digit central office code, and a 4-digit customer's number. In many parts of the country automatic toll switching systems⁴ are now in use by operators who complete more than 40 per cent of all toll calls by dialing directly to the called telephone in distant cities.

In order that customers may use these switching systems to dial their own toll calls, some automatic means for recording the necessary billing information on such calls must be provided.

PRESENT RECORDING AND CHARGING METHODS

Several types of automatic recording equipment are now in service in the Bell System.

Multi-unit registration (zone registration) has been in use for many years in a number of panel and No. 1 crossbar central offices whose rate structure permits bulk billing. This method can be used for calls which cost 6 or less message units for the initial period. Although zone registration is economical, it does not provide a detailed record of each call but merely scores the number of message units on a register associated with the customer's line.

Remote control zone registration has been serving customers in panel central offices since 1941. It is similar to multi-unit registration, but the timing and register control equipment is located in a tandem office instead of in each originating panel central office.

Automatic ticketing,⁵ which was developed some years ago for use in step-by-step central offices, does make a record of the details of each customer dialed call. A simple ticket printer is permanently associated with each outgoing trunk to produce an individual typewritten ticket for each call. Common relay equipment is used to furnish the called number, calling number, etc. to the printer. The information printed on the ticket is in detailed form and is similar to that prepared by the operator in manual operation. It can be used for billing the customer manually either on a detailed or a message unit basis.

A greatly improved form of recording, the Automatic Message Accounting⁶ (AMA) system, was introduced into the Bell System in 1948. In central offices having this equipment, all of the data required for billing of customer dialed calls are automatically perforated in code on paper tapes. These tapes are taken to an accounting center where they are processed by suitable machines to produce customers' bills. The recording machines are associated with the transmission circuits only when required to make a record, one recorder serving up to 100 such circuits. Recorders, together with their associated equipment, are installed in each central office arranged for local AMA recording. The information for each call is recorded on the tape in three stages, or entries. The initial entry is recorded after the customer has finished dialing. One time entry is recorded when conversation starts and another when conversation ends. For short-haul calls that are to be billed on a message unit basis, the initial entry contains only the calling office code and telephone number, and the charging rate. This information, together with the duration of the call, is sufficient for determining the charges. On toll calls which are to be billed in detail, the called office and telephone number are also

required. This system permits individual and two party customers in No. 1 and No. 5 crossbar offices to dial calls to telephones in their home area. In addition, customers in some No. 5 crossbar central offices may now dial directly to other areas. These facilities have been installed in No. 5 crossbar offices in Englewood, N. J., and in several other locations whose customers now may dial directly to about 13 million telephones in 13 metropolitan areas.

Relatively expensive recording equipment is required in each central office in the local AMA system. For new central offices this recording equipment is economical only if the toll and message unit calling rates are relatively high. The addition of local AMA recording equipment to existing offices is, in most cases, uneconomical.

CENTRALIZED AUTOMATIC MESSAGE ACCOUNTING

The Centralized Automatic Message Accounting system (CAMA) provides an economical means of recording billing data for customer dialed calls from many central offices that cannot justify local AMA. This system is economical because one group of recording equipment, located at a crossbar tandem office,⁷ can serve as many as 200 local central offices without requiring major changes in, or additions to, those offices.

The first crossbar tandem equipment arranged for CAMA was placed in service in Washington, D. C., in November, 1953. This equipment serves the customers in 85 central offices in Washington and in suburban Virginia and Maryland. They are able to dial each other directly and to dial their own calls to Baltimore and to other nearby toll points. Eventually, they will be able to dial their own calls to most points in the United States and Canada. Similar crossbar tandem CAMA equipments have been installed in Detroit, New York, San Francisco and Philadelphia.

The CAMA installation at Detroit enables the customers served by approximately 800,000 telephones in 99 Detroit panel and No. 1 crossbar local central offices to dial station-to-station multi-unit interzone and toll calls to 63 communities in Michigan and in nearby Canada. The map of Fig. 1 shows this dialing area.

THE CROSSBAR TANDEM SYSTEM

The crossbar tandem system into which CAMA has been introduced is used today in panel-crossbar and step-by-step areas. It receives calls from local dial central offices and completes them to other local central offices and to the toll network. It is also arranged to receive calls over

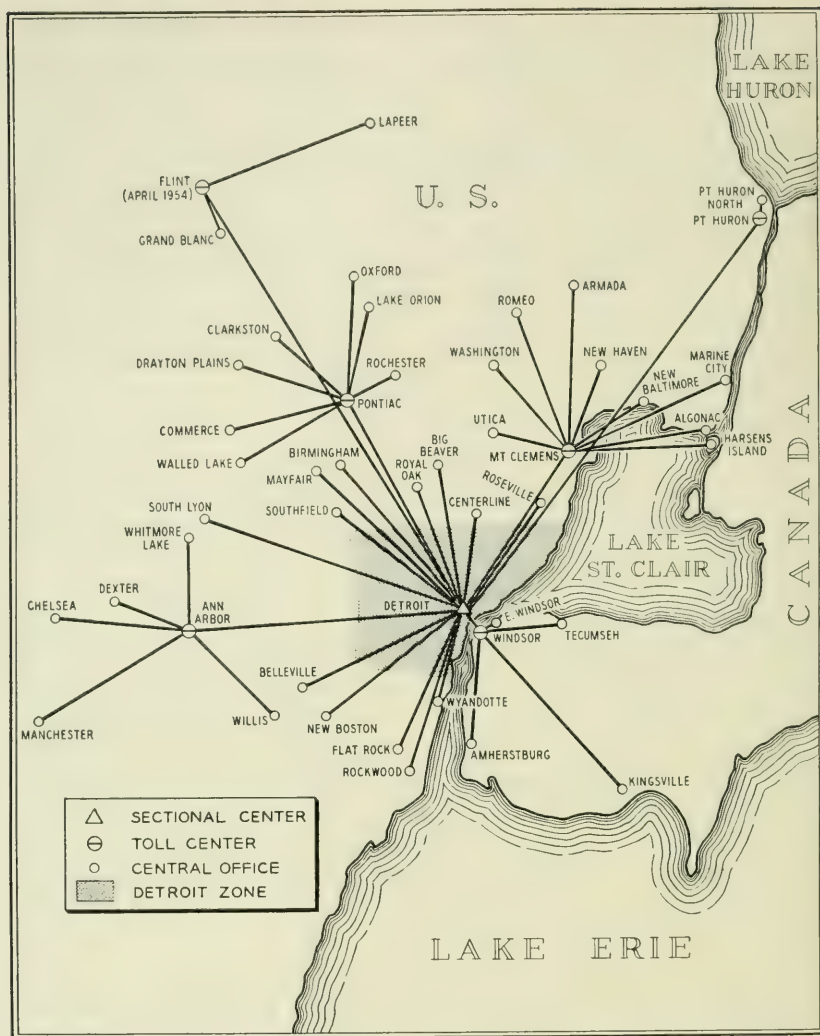


Fig. 1 — Points reached via Detroit CAMA at cutover in Dec., 1953.

intertoll trunks and complete them to other intertoll trunks or to local central offices. In many situations, it provides more efficient trunking facilities between central offices than do direct trunks and connects together offices with different signaling systems. Since the present crossbar tandem offices have in themselves no means for recording billing data, their use has been restricted to operator dialing, to customer dialing of

flat rate calls from all types of central offices, and to customer dialing of message unit and toll calls from central offices using one of the present methods of charging.

FIELD OF USE FOR THE CAMA SYSTEM

The CAMA system, as now developed, is suitable for use in panel-crossbar local areas. It is arranged to serve 7-digit calls only since facilities for 10-digit dialing are not available for panel and No. 1 crossbar central offices. Thus, in general, it can complete calls only to its own numbering area. However, provision is made for completing calls to one adjacent area, this area being selected by dialing "one-one" ahead of the listed 7-digit number.

BRIEF DESCRIPTION OF THE CAMA SYSTEM

If a customer makes a call that requires CAMA treatment, the call will be routed by the local central office to a crossbar tandem office arranged for CAMA recording. Until automatic means for identifying the calling customer's number for billing purposes is developed, an operator will be bridged on the connection at the tandem office to obtain the calling number and register it in the CAMA equipment by keying. The

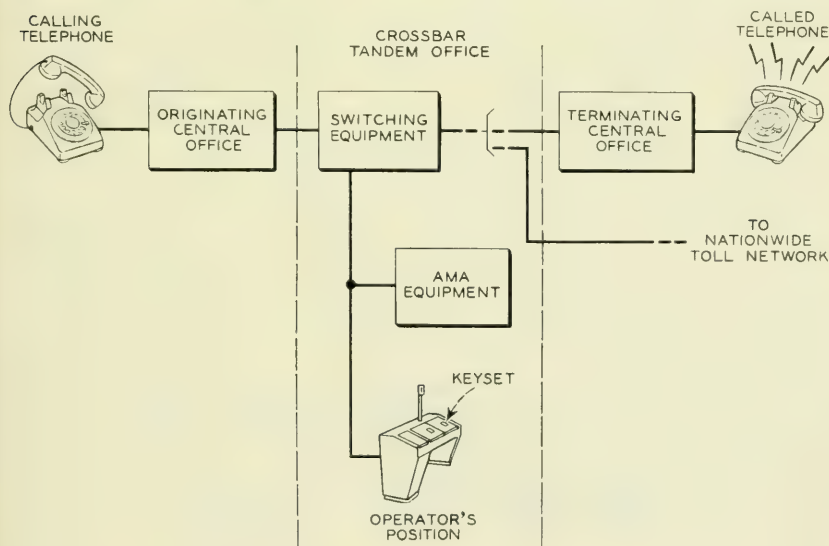


Fig. 2 - Simplified switching diagram.

information necessary for correctly charging for the call will then be recorded on paper tape by the automatic message accounting equipment located in the tandem office. This method of operation is shown in simplified form on Fig. 2.

A more detailed block diagram of the principal equipment units and their interconnections for a crossbar tandem CAMA office in a panel-crossbar area is shown in Fig. 3. Here the switching equipment consists of the conventional trunks, sender link frames, senders, markers and trunk link and office link frames for switching calls through the office. Such new units as the position link frames, positions, transverters, billing indexers, recorders, call identity indexers, and master timer constitute the major AMA equipments needed for recording the billing information for each call. Most of these have functions similar to corresponding local AMA equipments. AMA features have also been added to the trunks, sender links and senders.

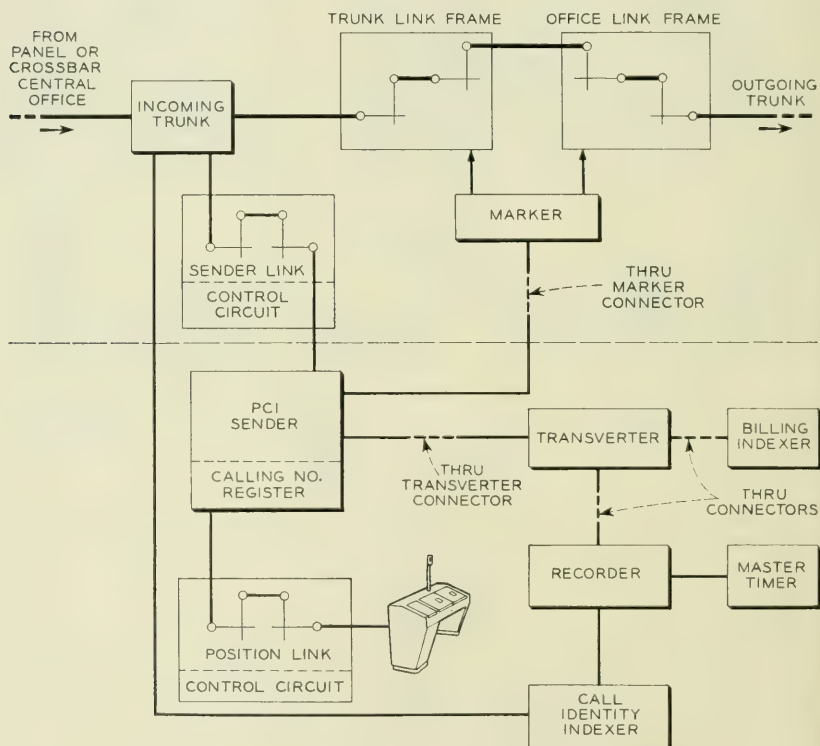


Fig. 3 — Block diagram of CAMA system in panel-crossbar areas.

FUNCTIONS OF THE CAMA SYSTEM

The functions of the system in establishing a connection and in recording the billing data may be divided into seven major groups as follows:

1. Operation of the sender link and control circuit in selecting an idle sender and connecting the selected sender to the incoming trunk.
2. Receiving and registering the called office code and number in the sender. The sender does not pulse the entire number forward until the AMA functions are completed.
3. Operation of the marker in establishing the connection through the switches and furnishing the sender with directions for completing the call.
4. Operation of the position link and control circuit in selecting an idle occupied position and connecting it to the calling customer.
5. Obtaining the number of the calling telephone verbally from the customer and keying it into the sender.
6. Connection of the sender to a transverter and billing indexer and the derivation of the billing data from the called and calling office codes and the rate class of the calling customer, and recording the charging information on the AMA tape.
7. Operation of the sender in transmitting information of the proper type to the terminating office, or if a toll call, to the next toll office in the chain.

Calls from Panel and Crossbar Customers

The first three functions in a crossbar tandem CAMA-equipped office in a panel-crossbar area are the same as in a non-CAMA office. Since published information on these features⁷ is available, they will not be described in detail.

The fourth major function is handled by the position link and control circuit which is shown in block diagram form in Fig. 4. This circuit consists of primary and secondary crossbar switch links and control circuits in duplicate. Each group functions independently to serve calls to the same 40 senders and can connect to two different groups of 50 positions. In case of failure of one link group, the other link group will continue to serve calls to the 40 senders. The control circuits are arranged in such a manner that all senders and all positions receive essentially equal treatment.

When the position link connects the sender to an idle CAMA position, the operator obtains the calling number verbally from the customer and

records that number in the sender by the operation of numerical keys located at her position. The CAMA switchboard shown in Fig. 5 is a cordless board of modern sheet metal design.

The sixth group of functions — that of converting the data into the desired form and perforating it on the paper tape — is performed jointly by the transverter, the billing indexer, the recorder and the call identity indexer. The transverter is very similar to that used in local AMA. It registers the calling and called number received from the sender, registers the billing data received from the billing indexer and controls the recorder in the perforation of the initial entry.

The billing indexer is strictly a translating circuit. It receives the calling and called office codes and the customer rate class from the transverter and converts this information into a form which the transverter and recorder can use. It provides a 1-digit billing index, which denotes the charging plan to be used, and provides a type of initial entry indica-

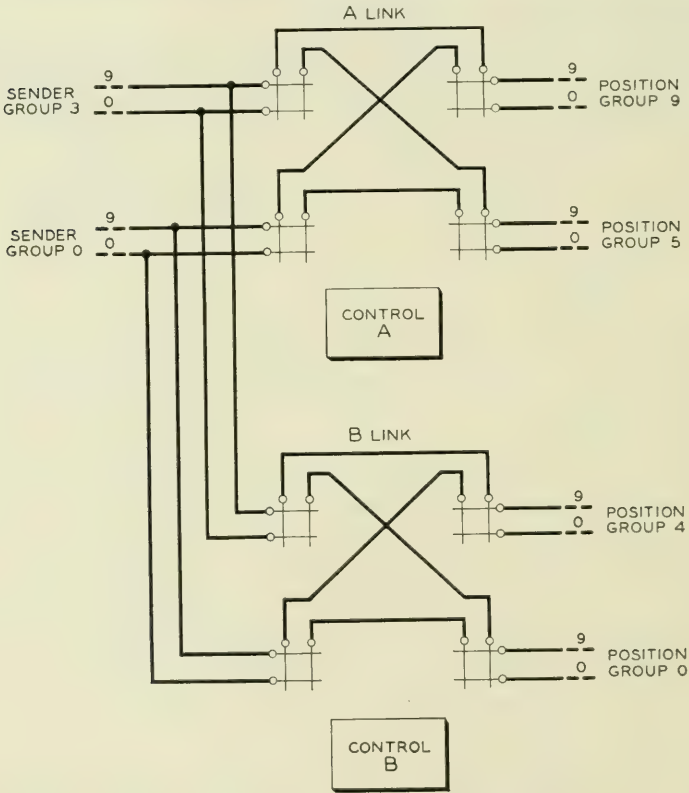


Fig. 4 — Position link and control circuit.

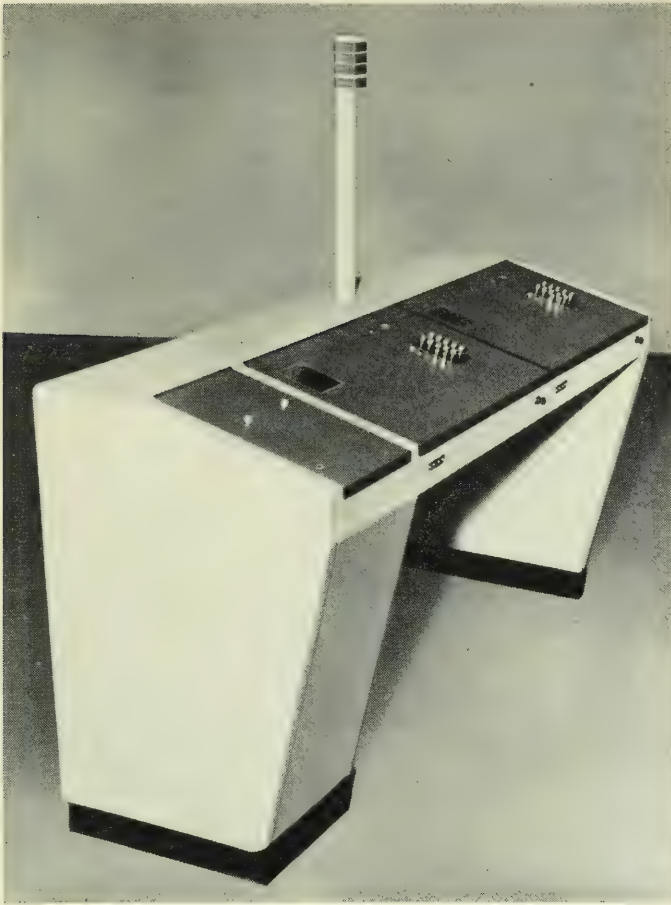


Fig. 5 — Switchboard.

tion which tells the transverter whether to perforate a two- or a four-line initial entry. The two-line initial entry used for calls billed on a message unit basis contains only the calling office code and telephone number, the billing index and the trunk identity whose function is discussed later. This information, together with the duration of the call, is sufficient for billing. Four-line entries contain, in addition, the called office code and telephone number. They are used for detail billed toll calls and for those bulk billed calls on which all details of the call are required for record purposes. Fig. 6 shows a simplified schematic of the billing indexer. The calling office code combined with the customer rate class, chooses a particular rate treatment relay. This rate treatment is common to all cus-

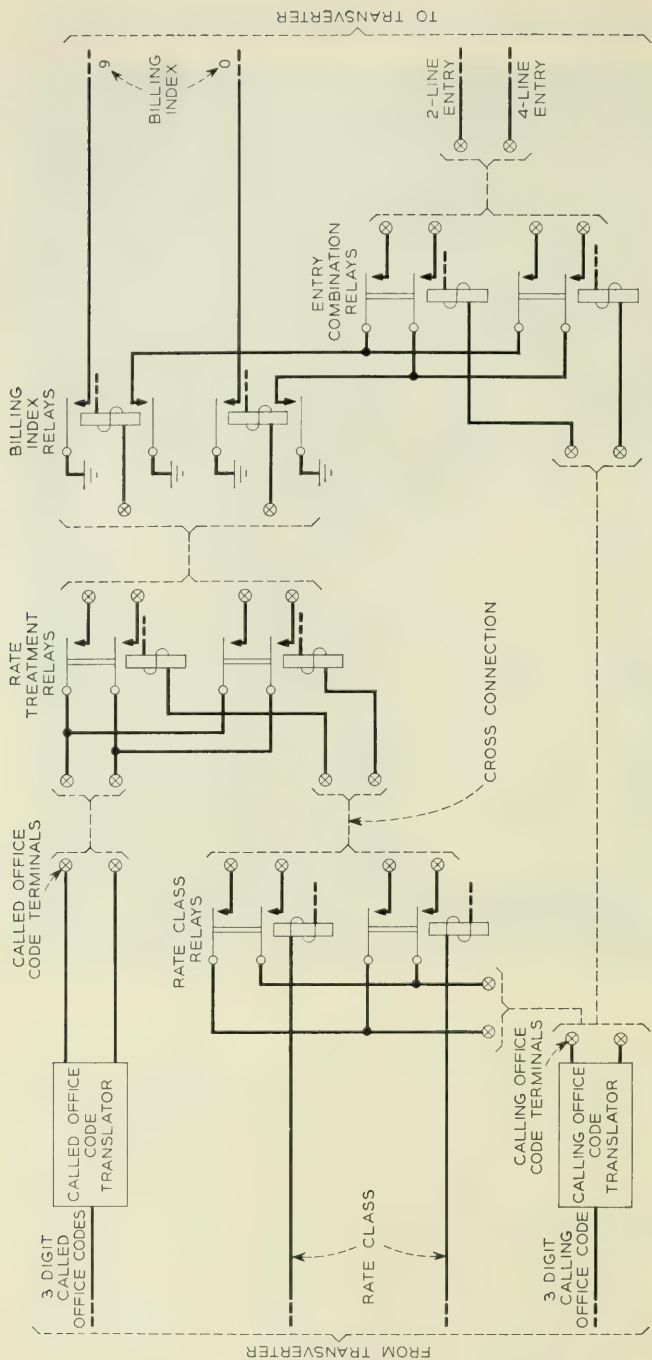


Fig. 6 — Simplified schematic of billing indexer.

tomers in the area who are charged alike for their calls through the tandem office. To actually determine the billing index or charge treatment for each call, the rate treatment is modified by the called office code by means of the rate treatment relays. Since calls with the same billing index may require a 2-line entry if originated in some offices and 4-line entry if originated in other offices, the billing index and calling office code information are translated jointly by means of the entry combination relays to produce either the 2-line or 4-line indication.

The recorder,⁸ call identity indexer and master timer circuits are of the same type as used in local AMA and perform the same functions. The recorder perforates initial entries as directed by the transverter. It also perforates a timing entry at the beginning of conversation and another at the end of conversation. The call identity indexer, one of which is associated with each recorder, identifies the trunk used on a call as a particular one of the maximum 100 served by a recorder. This enables the recorder to perforate that identity on initial and timing entries. The identity is used by the accounting center to gather together the three entries involved on each call.

The master timer⁹ keeps the recorders continually informed as to the correct time.

When the initial entry is completely recorded on the AMA tape, the sender completes its task of pulsing the called number forward and then releases. The transmission path is now completed through the tandem office and the only CAMA functions remaining are the perforations of the timing entries mentioned above.

ACCOUNTING CENTER PROCESS

The accounting center process for CAMA is the same as for local AMA. It automatically assembles the three bits of information pertaining to each call, computes the conversation time on all calls, sorts by the type of call, prices each call either in terms of message units for bulk-billed calls or in terms of dollars for detail billed toll calls and brings together the records of all calls made by each customer.

MAINTENANCE FEATURES

To properly maintain the AMA recording facilities, test circuits are provided for testing the major features of the CAMA equipment. An automatic incoming trunk test circuit tests the CAMA trunks. An automatic sender test circuit tests the CAMA senders in much the same way that the present sender test circuit tests non-AMA senders. Facilities are also provided for making operating tests of the position links, posi-

tions, transverters and billing indexers. As in the local AMA system, testing of the recorder features is done by the test unit on the master timer frame.

The AMA circuits are provided with many self-checking features which detect trouble while a call is being handled. When a trouble is detected, the AMA circuits connect momentarily to a recording circuit called a "trouble indicator" which lights lamps to indicate how far the call has progressed and which of the common control circuits were used on the call. This information aids the maintenance force in locating the trouble.

FUTURE DEVELOPMENTS

As stated earlier, use of Centralized Automatic Message Accounting by panel and crossbar customers will be restricted initially to calls to the home area and one foreign area. The centralized recording will be done initially at crossbar tandem offices with operators identifying the calling telephones. Ultimately, customers served by all types of dial local central offices will be able to dial their own calls — local or nationwide. Operator identification of individual and two-party lines will be replaced in many cases by automatic identification. The centralized recording equipment will be located in various types of tandem and toll offices as determined by the economics of each case.

CONCLUSION

The development of Centralized Automatic Message Accounting arrangements is another major step toward nationwide customer dialing from central offices which cannot be economically equipped with local AMA recording equipment.

REFERENCES

1. Pilliod, J. J., Fundamental Plans for Toll Telephone Plant, B.S.T.J., **31**, pp. 832-850, 1952.
2. Clark, A. B., and Osborne, H. S., Automatic Switching for Nationwide Telephone Service, B.S.T.J., **31**, pp. 823-831, 1952.
3. Nunn, W. H., Nationwide Numbering Plan, B.S.T.J., **31**, pp. 851-859, 1952.
4. Shipley, F. F., Automatic Toll Switching Systems, B.S.T.J., **31**, pp. 860-882, 1952.
5. Friend, O. A., Automatic Ticketing of Telephone Calls, A.I.E.E. Trans., **63**, pp. 81-88, 1944.
6. Meszar, J., Fundamentals of the AMA System, A.I.E.E. Trans., **67**, Part I, pp. 255-269, 1950.
7. Collis, R. E., Crossbar Tandem System, A.I.E.E. Trans., **69**, Part II, pp. 997-1004, 1950.
8. Cahill, H. D., Recording on AMA Tape in Central Offices, Bell Labs. Record, **29**, p. 565, Dec., 1951.
9. Jordan, W. C., The AMA Timer, Bell Labs. Record, **30**, p. 122, March, 1952.

The Wave Picture of Microwave Tubes

By J. R. PIERCE

(Manuscript received March 12, 1954)

Many microwave tubes make use of a long electron beam. The radio frequency excitation on such a beam can be expressed in terms of two space-charge waves, one of which has negative energy and negative power flow. The electron beam may pass through resonators, through lossy surroundings, through slow-wave circuits. In this paper the low-level operation of klystrons, resistive-wall amplifiers, casitrons, space-charge-wave amplifiers, traveling-wave tubes and double-stream amplifiers is explained in terms of waves on electron beams and on circuits. Noise is discussed in terms of such waves.

INTRODUCTION

There are many different ways in which one can make a valid analysis of the low-level or small-signal behavior of the many types of microwave tubes which use long electron beams. Which way one should choose depends partly on one's purpose in making the analysis, and partly on the particular problem to be solved.

All of these analyses lead at some point to waves or modes of propagation: waves which travel along an electron stream, along a circuit, or along the two together; waves which are unattenuated or which increase or decrease with distance. Sometimes, the analysis starts out with electron current, electron velocity and circuit dimensions as the fundamental physical quantities, just as network analysis can start out with inductance, capacitance and resistance. However, an analysis can start out instead with waves, their propagation constants and their characteristic impedances as the fundamental physical bases of the analysis. We might argue that as we are to end with waves, we may well start with waves. As it turns out, the picture of the operation of various tubes in terms of waves is simple and pleasing.

It is the purpose of this paper to present a picture of the operation of microwave tubes in terms of waves. This may be of some interest to those outside of the tube field, in that it gives an account of many recent devices. For experts in the field it can serve as an introduction to a method of analysis which is fairly recent and which may be unfamiliar.

In this analysis, certain simplifications are made. One underlying simplification is that of linearity; it is assumed that at low signal levels the behavior of the electron stream, which is inherently non-linear, can be represented by that of a truly linear system.

As this paper purports to give an accurate and useful picture of the low-level operation of microwave beam devices rather than an exhaustive discussion, some details have been omitted or passed over lightly because they seemed to be of secondary importance. Material which may be unfamiliar to workers in other fields but which is important as background is presented in appendices A to C. Various points can be pursued further in the literature. References to publications and to those responsible for various advances are not given in the body of the paper or in the appendices; they are given for each topic in Appendix D.

SPACE-CHARGE WAVES

Many microwave tubes embody a long, narrow electron beam surrounded by a conducting tube and focused or confined by a longitudinal magnetic field. At low levels of operation, the radio-frequency disturbances on such an electron stream can be expressed in terms of space-charge waves.

In these waves, two forms of energy are of primary importance: electrostatic energy associated with the bunching together of electrons, and kinetic energy, associated with differences in the velocities of the electrons. Thus, the waves may be called electromechanical; the electric energy which we associate with waves in transmission lines and waveguides is present, but the magnetic energy is replaced by kinetic energy. In circuit terms, we have an electrical capacitive element, but the inductive element is inertial, not magnetic in nature. When the electron and wave velocities are slow compared with the velocity of light, the magnetic fields produced by the electron convection current are negligible.

There may be many space-charge modes or waves in an electron stream, some with complex radial and angular variations of amplitude over the electron stream. Two waves predominate in the operation of tubes, however, and one simplification we will make is to deal with these only, and to disregard other modes of propagation on the electron stream. Appendix A discusses such a pair of waves in a simplified physical system.

We can associate with these two waves an ac electron convection current i and an ac electron velocity v . Either we can assume that the electron beam is narrow and disregard the fact that these quantities vary

across the beam, or we can deal with peak or effective values much as in the case of voltages and currents in waveguides.

These ac quantities are assumed to contain a factor

$$e^{j\omega t} e^{-j\beta z}$$

That is, they vary sinusoidally with time and with distance, and (assuming β to be positive) propagate in the $+z$ direction. The phase constants β of the two waves will be called β_1 and β_2 . For beams of moderate charge they are very nearly

$$\beta_1 = \frac{\omega}{u_0} + \frac{\omega_q}{u_0}$$

$$\beta_2 = \frac{\omega}{u_0} - \frac{\omega_q}{u_0}$$

Here u_0 is the electron speed, ω is the operating radian frequency and ω_q is the effective plasma radian frequency.

The plasma frequency of the electron beam ω_p is given by

$$\omega_p^2 = \frac{\frac{e}{m} \rho_0}{\epsilon}$$

Here e/m is the charge-to-mass ratio of the electron, ρ_0 is the charge density and ϵ is the dielectric constant of vacuum. In terms of ω_p , ω_q may be expressed

$$\omega_q = R\omega_p$$

Here R is a factor somewhat less than unity which depends on the geometry of the electron beam, on ω and ω_p , and on the velocity distribution of the electrons (see Appendices A and C).

Let us consider the simple case in which R is unity and the effective plasma frequency is equal to the plasma frequency. The phase velocities v_1 and v_2 of the two waves, which are ω divided by β , are

$$v_1 = \frac{u_0}{1 + \frac{\omega_p}{\omega}}$$

$$v_2 = \frac{u_0}{1 - \frac{\omega_p}{\omega}}$$

Thus, the first wave has a phase velocity less than that of the electrons; it is a slow wave, and the second wave is a fast wave.

Suppose we make up a radio-frequency pulse out of various frequency components of one wave. The pulse envelope generally travels with a different velocity from that of the rf sinusoids under the envelope. The velocity of the envelope is called the group velocity. The group velocity is the velocity with which a signal is transmitted. The direction of the group velocity is the direction in which causality acts (for some waves the phase velocity and the group velocity have opposite directions). The group velocity tells in which direction energy flows, and the power flow P is the stored energy per unit length, W , times the group velocity, v_g .

$$P = Wv_g$$

The group velocity is given by

$$v_g = \frac{1}{\partial\beta/\partial\omega}$$

We see that for our assumption ω_q is equal to ω_p , the group velocity for each wave or mode is u_0 , the velocity of the electrons in the beam

$$v_g = u_0$$

Thus, of the two waves, the first has a phase velocity slower than that of the electrons, the second has a phase velocity faster than that of the electrons, and each has a group velocity equal to that of the electrons.

A simple discussion of power flow is given in Appendix B. In describing the excitation of the electron stream we can use the convection current i together with a quantity U which is analogous to voltage. In terms of the ac electron velocity v ,

$$U = -\frac{u_0}{|e|} v$$

The real power flow P is given by

$$P = \frac{1}{2}(iU^* + i^*U)$$

This relation is justified in Appendix B.

For each of the two waves the voltage U bears a constant ratio to the current i ; this ratio is the characteristic impedance K of the wave. We find that

$$K_1 = \frac{U_1}{i_1} = -2\frac{\omega_q}{\omega} \frac{V_0}{I_0}$$

$$K_2 = \frac{U_2}{i_2} = 2\frac{\omega_q}{\omega} \frac{V_0}{I_0}$$

Here V_0 is the accelerating voltage specifying the electron velocity u_0 and I_0 is the total beam current.

We see that the characteristic impedance K_1 of the slow wave is negative. This means that the power flow in the $+z$ direction is negative. We could also say that positive power flows in the $-z$ direction, but this may carry an unfortunate implication as to the direction in which causality acts. An example may be helpful.

Fig. 1 shows an electron beam acted on by the fields of two devices A and B . The fields in A are such as to set up the slow wave only. This travels between A and B . The fields of B are such as to just remove the slow wave entirely, so that the electron beam leaves B with no ac disturbance on it. The electron velocity u_0 , phase velocity v , group velocity v_g and negative power flow $-P$ are all directed in the $+z$ direction, that is, to the right.

We must remove a power P from A to set up the slow wave. A power $-P$ flows from A to B . We must add a power P to B to remove the slow wave from the electron beam. Causality acts from A to B . To change the amplitude of the slow wave between A and B we must change the fields in A , not the fields in B .

The power flow is the group velocity times the stored energy per unit length. As the group velocity for the slow wave is positive and the power flow is negative, we see that the stored energy must be negative.

If we moved with the electrons and observed the waves, we would find that the average kinetic energy associated with the ac electron velocity was equal to the average potential energy of the electric field, and that both were positive; this is characteristic of waves in a stationary medium. The kinetic energy of the electrons relative to a fixed observer is proportional to the square of their total velocity, that is, the ac velocity plus the average velocity. The average velocity is larger than the ac velocity, so that energy terms involving the product of the average

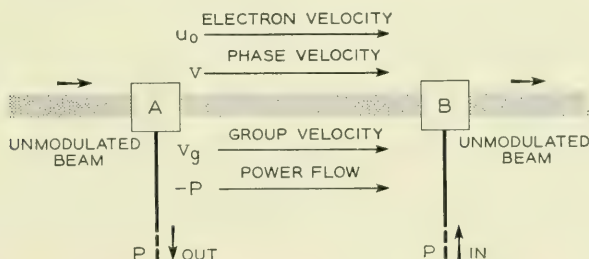


Fig. 1 — Device A sets up the slow space-charge wave only, and device B removes it. u_0 , v , v_g and $-P$ are respectively the electron velocity, the phase velocity, the group velocity and the power flow between A and B .

velocity and the ac velocity are larger than terms involving the square of the ac velocity. The product terms may be negative or positive.

We can understand the negative energy of the slow wave qualitatively through a simple argument of a somewhat different sort. In the slow wave, the charge density is greatest in regions of less-than-average velocity and least in regions of more-than-average velocity, so that the electron beam has less total kinetic energy in the presence of the slow wave than it does in the absence of the slow wave. How does this come to be? Suppose that we move with the wave; we then see electrons moving in an electric field which is constant with time, and hence, as electrons move through the field their velocities vary as the square root of the potential. Relative to the wave, the electrons move slowest in the low-potential regions, and correspondingly, they are bunched together in regions of low potential. Now, for the slow wave the total electron velocity is the arithmetic sum of the wave velocity and the electron velocity relative to the wave, so if the electrons are bunched in regions of lowest velocity relative to the wave they are necessarily bunched in the regions of least total electron velocity, and the kinetic energy of the slow wave is thus negative.

In the case of the fast wave, the electrons travel backward relative to the wave. The total electron velocity is the arithmetic difference between the wave velocity and the electron velocity relative to the wave. Hence, the total electron velocity is greatest at the bunches, where the velocity relative to the wave is least, and the kinetic energy of the fast wave is positive.

THE KLYSTRON

We can explain the operation of a number of types of vacuum tubes in terms of space-charge waves. Consider the klystron, illustrated in Fig. 2. The voltage produced across the input resonator by the input signal sets up on the electron beam both the slow and the fast space-charge waves in equal magnitudes and so phased that the velocities v , or the voltages U add, while the currents cancel. Thus, just beyond the input resonator, the beam has an ac velocity; it is velocity modulated, but it has no ac convection current.

Because the two space-charge waves, one with negative power flow and one with positive power flow, are set up with equal magnitudes, the ac power flow in the beam between the input and the output resonators is zero. The input resonator neither adds power to nor subtracts power from the beam.

Because the two waves have different phase velocities, their relative phase changes as they travel along the beam. If we go along the beam a

distance L such that

$$2 \frac{\omega_q}{u_0} L = \pi$$

we will find that the ac velocities of the two waves cancel and their currents add. If at this point we put an output resonator, the current will produce a voltage across the resonator which will act on the electron beam to set up new components of the slow and the fast waves.

If the resonator is on tune, so that it acts as a resistive impedance, the phase of the voltage is such with respect to the space-charge wave producing it that the new component of the fast space-charge wave

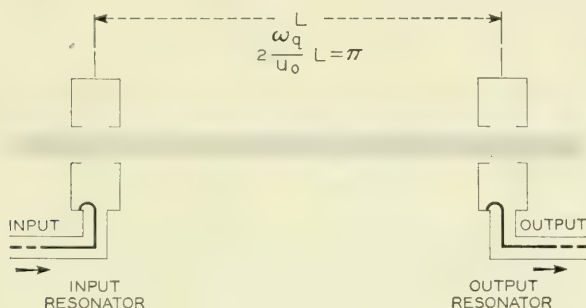


Fig. 2 — In a klystron the input resonator sets up slow and fast space-charge waves so phased that the velocities add and the currents cancel. At the output resonator the currents add and the velocities cancel. The voltage across the output resonator increases the amplitude of the slow, negative-power wave and decreases the amplitude of the fast, positive-power wave.

subtracts from the old component, while the new component of the slow space-charge wave adds to the old component. Thus, while to the left of the output resonator the two space-charge waves have equal magnitudes, so that the net power flow is zero, to the right of the output resonator the slow space-charge wave has a greater magnitude than the fast space-charge wave, so that the power flow in the beam is negative. The missing power appears as the output from the output resonator.

Of course, klystrons are frequently used in the nonlinear range of operation, and the distance L between resonators may be chosen differently from other considerations.

THE RESISTIVE-WALL AMPLIFIER

Consider a tube much like a klystron, but in which the electron beam is surrounded by a glass tube coated with lossy material, such as graphite, as shown in Fig. 3.

As in the klystron, the input resonator produces both the slow and the fast waves with equal magnitudes. As each wave travels, it induces currents in the resistive wall surrounding it and dissipates power in the wall. Thus, the power in each wave must decrease as the wave travels.

The fast wave has a positive power, and so for its power to decrease the amplitude must decrease. Thus, in the resistive-wall region the amplitude of the fast space-charge wave decreases exponentially with distance.

Because the slow space-charge wave has a negative power, its power can decrease only if the amplitude of the wave increases, so that the power flow becomes less (more negative). Thus, in the resistive-wall region the amplitude of the slow space-charge wave increases exponentially with distance; the wave has a negative attenuation; it is amplified as it travels.

If we put the output resonator far from the input resonator, the amplitude of the fast space-charge wave will be very small there, but the amplitude of the slow space-charge wave may be very large. Its current will produce a large voltage across the output resonator. As in the case of the klystron, this voltage will increase the amplitude of the slow space-charge wave, thus decreasing the power flow in the electron stream.

The resistive wall amplifier has a feature which the klystron lacks; the process of amplification involves an actual growing wave along the electron stream.

THE EASITRON; INCREASING WAVE IN A LOSSLESS SYSTEM

Consider a tube somewhat similar to the resistive wall amplifier, but in which the beam is surrounded, not by a lossy tube, but by a series of pill-box resonators, as shown in Fig. 4. Imagine that the resonators are so tuned that at the operating frequency they present a lossless negative susceptance to the electron beam.

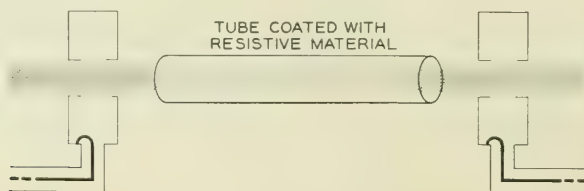


Fig. 3 — In a resistive-wall amplifier the currents excited in the lossy wall by the slow, negative-power wave decrease the power in the wave, so that the amplitude of the wave must increase.

The impedance an electron beam sees in traveling through free space or in a concentric lossless tube is capacitive. In section 1 the space-charge waves were described as involving the stored energy of the electric field, capacitive in nature, and the kinetic energy of the electrons, which has an inductive effect. We might liken the beam and its capacitive circuit to the ladder network of Fig. 5. We know that such a network supports waves.

When the charge of the beam sees a negative susceptance, the behavior is much as if the capacitances in the ladder network of Fig. 5 were negative.* In this case the waves characteristic of the circuit are not traveling waves, but are a pair of waves, one of which decays with distance and one of which increases with distance. Neither has any net stored energy.

We can express the propagation constants of the waves much as in the section on, "Space-Charge Waves," but the effective plasma frequency ω_q is now imaginary; we will call it $j\omega_q'$. The phase constants β_1 and β_2

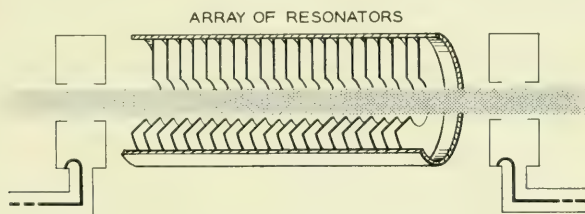


Fig. 4 — In the easitron, resonators surrounding the beam change the susceptance the electrons see from positive to negative. The system no longer supports two traveling waves, but rather, a growing and a decaying wave.

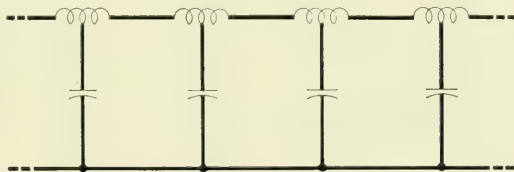


Fig. 5 — If the capacitances in this ladder network were negative it would support growing and decaying waves rather than traveling waves.

* Some care must be used in arriving at proper equivalent circuits. For instance, neither of the electric waves on a ladder network has negative energy if the network is set in motion, but we have seen that one of the longitudinal space-charge waves does have negative energy. If both the capacitances and the inductances of a ladder network are negative, the waves on the network will have negative energies.

become

$$\beta_1 = \frac{\omega}{\omega_0} + j \frac{\omega_q'}{\omega_0}$$

$$\beta_2 = \frac{\omega}{\omega_0} + j \frac{\omega_q'}{\omega_0}$$

The characteristic impedances become

$$K_1 = -j2 \frac{\omega_q'}{\omega} \frac{V_0}{I_0}$$

$$K_2 = j2 \frac{\omega_q'}{\omega} \frac{V_0}{I_0}$$

The fact that the characteristic impedances of the two waves are imaginary means that neither of the waves alone has any power flow. Neither of the waves can very well carry power. The amplitudes change with distance; hence for each wave $U i^*$ and $i U^*$ increase or decrease with distance. But, the circuit and the electron beam are lossless, and the power cannot change with distance. As the waves do have a group velocity, neither has any stored energy. Does this mean that the beam cannot carry any power? The beam can carry power, just as a filter in its stop band can carry some power from a source at one end to a resistive load at the other end. The power flow is still given properly in terms of the total current i and the total voltage U by the same expression used in section 1. Suppose that the two waves have currents i_1 and i_2 . Then the total power flow is

$$P = \frac{1}{2}[(i_1 + i_2)(K_1^* i_1^* + K_2^* i_2^*) + (i_1^* + i_2^*)(K_1 i_1 + K_2 i_2)]$$

$$P = \frac{1}{2}[(K_1 + K_1^*)(i_1 i_1^*) + (K_2 + K_2^*) i_2 i_2^* + i_1 i_2^* (K_1 + K_2^*) + (i_1 i_2^* (K_1 + K_2^*))^*]$$

First consider the case in which ω_q is real and for which the characteristic impedances are real and

$$K_1 = -K_2$$

In this case

$$P = K_1 i_1 i_1^* + K_2 i_2 i_2^*$$

This is the familiar case of unattenuated waves. The total power is the power of each wave calculated individually.

Let us now consider the case in which the effective plasma frequency

is imaginary. In this case we can write

$$K_1 = -jK_0$$

$$K_2 = +jK_0$$

where K_0 is real. We have

$$P = [-ji_1i_2^*K_0 + (-ji_1i_2^*K_0)^*]$$

Either wave alone carries no power; there is power flow only when the two waves are present simultaneously. The two waves vary with distance as

$$e^{-j(\omega/u_0)z} e^{(\omega_{q'}/u_0)z}, \quad e^{-j(\omega/u_0)z} e^{-(\omega_{q'}/u_0)z}$$

so the $i_1i_2^*$ is constant with distance. If this were not so the power would change with distance, but as the resonators have been assumed to be lossless, neither taking power from the beam nor adding power to the beam, this is impossible. Thus, in a lossless system an increasing wave is always one of a pair, and the other member decreases with distance in such a way as to keep the product of the amplitudes of the two waves constant with distance. Neither the increasing wave alone nor the decreasing wave alone carries any power, but the two together can carry power.

We will note that in the easitron the direction of the group velocity, that is, the direction of causality, is the direction of electron flow. Thus, the waves are both set up at the input resonator; it is there that boundary conditions on both current and voltage must be satisfied.

COUPLING OF MODES OF PROPAGATION

We know that waves which increase and decrease exponentially with distance are characteristic of a ladder network in which the susceptances of the shunt and series arms have the same signs. They occur in other networks as well. Consider a smooth transmission line loaded periodically with shunt capacitances, as shown in Fig. 6. Each capacitance reflects



Fig. 6 — Capacitances connected across a smooth transmission line periodically couple the forward and backward waves and produce stop bands characterized by growing and decaying waves.

part of a wave approaching it. In other terms, each capacitance acts to couple one mode or wave (say the forward wave) to another (say, the backward wave). When the distance between the capacitances is such that the couplings reinforce, that is, near a half wavelength in this case, the system is a filter in its stop band; it does not transmit traveling waves, but supports rather a wave which increases exponentially with distance and a wave which decays exponentially with distance. Neither of these waves alone carries any power.

The space-charge waves of an electron stream can be coupled to one another, to a space-charge wave of another stream, or to an electromagnetic wave. In any of these cases we can have increasing waves.

THE SPACE-CHARGE-WAVE AMPLIFIER

Consider an electron beam surrounded by a series of metallic tubes $A, B, A, B \dots$, alternately at different potentials with respect to the cathode from which the electrons come, as shown in Fig. 7. The impedances of the space-charge waves will be different in tubes A from what they are in tubes B . The behavior of this system is much like that for the transmission line system shown in Fig. 8, in which we have alternate line sections of different characteristic impedances K_A and K_B . We know that such a series of line sections forms a filter with stop bands.

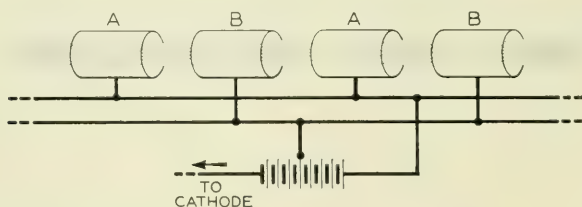


Fig. 7 — The impedances of waves in an electron beam passing through electrodes at alternately higher and lower potentials differ in regions of different potentials. This can result in stop bands characterized by growing and decaying waves. Such a device is a space-charge-wave amplifier.

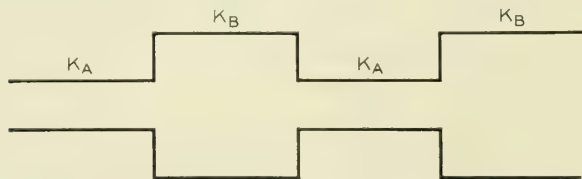


Fig. 8 — A transmission line with alternating sections of impedances K_A and K_B is somewhat analogous to the space-charge-wave amplifier.

In the case of the space-charge-wave structure of Fig. 7, the stop band occurs for conditions near that in which for both sections A and B the section lengths L_A and L_B are such that

$$2 \frac{\omega_{qA}}{u_0} L_A = \pi$$

$$2 \frac{\omega_{qB}}{u_0} L_B = \pi$$

Here ω_{qA} and ω_{qB} are the effective plasma frequencies for sections A and B .

A structure such as that of Fig. 7 can be interposed between input and output circuits, such as resonant cavities, to give a space-charge-wave amplifier dependent for its action on the growing wave of the pair.

THE TRAVELING-WAVE TUBE

In the space-charge-wave tube, the two waves which are coupled together have different velocities, just as the forward and backward waves on an electron stream have different velocities. Hence, they can be coupled strongly only through the use of some periodic structure in which the period is related to the difference in phase constants of the two waves.

In a traveling-wave tube we can have coupling between a space-charge wave and a wave traveling on a circuit, and both of the waves can have velocities which are nearly or exactly the same.

Here we must consider two different cases. If both of the coupled waves carry power in the same direction (that is, if the power is positive for both, or negative for both), coupling cannot result in a stop band, but only in transfer of power between one wave and the other. In order to have a stop band, power which we try to send in on one wave must come back to us on the other. Hence, to produce a stop band and gaining waves, the two coupled waves must carry powers with opposite signs.

A traveling-wave tube can consist of a helix of wire, which can support a slow electromagnetic wave, surrounding an electron beam, as shown in Fig. 9.

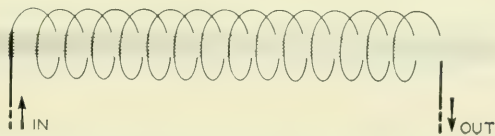


Fig. 9 — The vital elements of a traveling-wave amplifier are an electron stream and a slow-wave circuit which may be a helix surrounding the electron stream.

Traveling-wave tubes really involve at least four waves: two space-charge waves and two circuit waves. Usually, the backward circuit wave is so far out of synchronism with the space-charge waves that we can neglect its coupling with them. Further, if the space-charge waves are well separated in velocity, that is, when ω_d is large enough, then when one is coupled to the circuit wave the other isn't, and so we can get some idea of traveling-wave tube operation by considering waves in pairs. The simple mathematics of such coupling is given in Appendix D.

In Fig. 10, the behavior of various phase constants, plotted as a function of ω/u_0 , is shown qualitatively. Here ω is radian frequency and u_0 is electron velocity. We may consider that ω/u_0 is varied by changing the electron velocity u_0 and keeping the frequency ω constant. The horizontal line β_c is the phase constant of the forward circuit wave in the absence of electrons, or when the coupling to the electrons is zero. β_s and β_f are the phase constants of the slow and fast space-charge waves, respectively, with zero coupling to the circuit wave. For the slow space-charge wave, the power flow is negative, while for the circuit wave and the fast space-charge wave the power flow is positive. Thus, for coupling between the slow

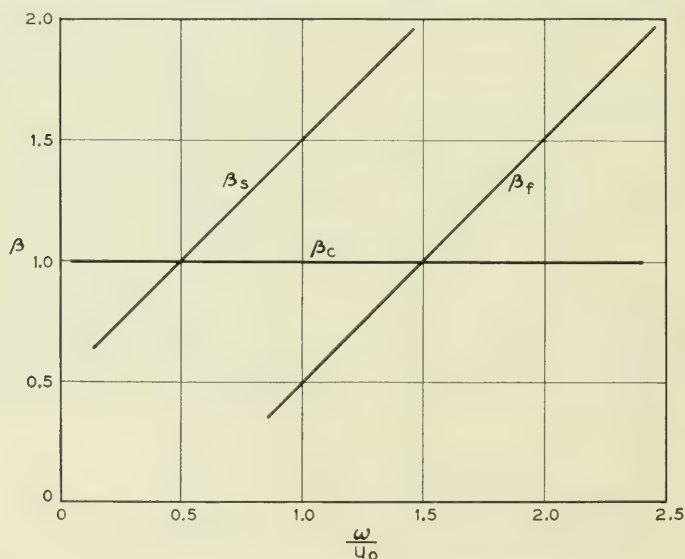


Fig. 10 — Suppose that at a constant radian frequency ω we change the electron velocity u_0 in a traveling-wave tube. If the waves of the electron stream were not coupled to the waves of the helix, the phase constants, β_c of the forward circuit wave, β_s of the slow wave, and β_f of the fast wave, would vary approximately as shown.

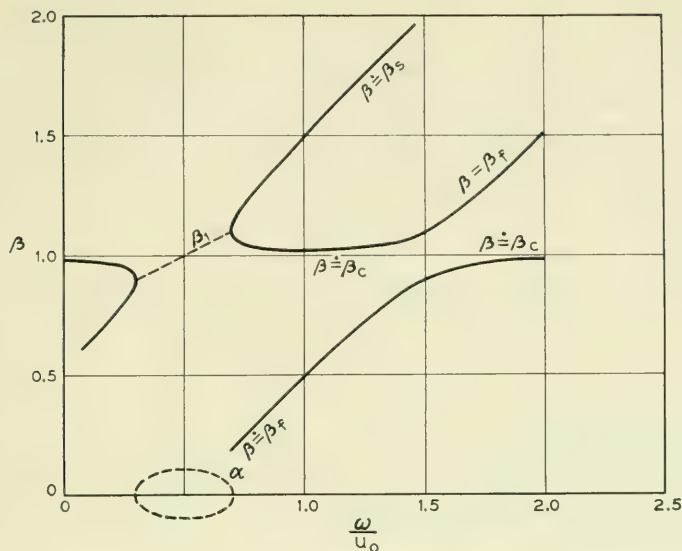


Fig. 11 — Because of coupling of the space-charge waves to the forward circuit wave, gain is produced near $\beta_c = \beta_s$, while the curves sheer off from one another near $\beta_c = \beta_f$.

space-charge wave and the circuit wave we can have a stop band, while for coupling between the circuit wave and the fast space-charge wave we cannot.

The consequences of the couplings between the circuit wave and the space-charge waves near the intersections of β_c with β_s and β_f are illustrated in Fig. 11.

We see from Fig. 11 that near synchronism between the circuit wave and the fast space-charge wave ($\beta_c = \beta_f$ for no coupling) these waves combine so that for any given value of ω/u_0 there are always two distinct real values of β . This is typical for coupling between modes with power flows of the same sign. At $\beta_c = \beta_f$ each of the two mixed waves has equal energies in the circuit and in the electron stream.

Near synchronism between the circuit wave and the slow space-charge wave ($\beta_c = \beta_s$ in absence of coupling) these two waves combine so that over a range of ω/u_0 near $\beta_c = \beta_s$, β has two complex values with the same real part and with equal and opposite imaginary parts. We can write this as

$$\beta = \beta_1 \pm j\alpha; \quad -j\beta = -j\beta_1 \mp \alpha$$

This corresponds to an attenuated and a growing wave with the same phase velocity. In Fig. 11, β_1 and α are plotted as dashed lines.

Over the range of ω/u_0 for which the waves are attenuated ($\alpha \neq 0$) the net power flow in each of the modes is zero. The power flow in the electron stream is equal and opposite to that in the circuit. Such behavior is characteristic when two modes with power flows of opposite signs are coupled. It is characteristic of the stop band of an electric wave filter.

The curves of Fig. 11 exhibit the same behavior that has been found by other means, although similar curves are sometimes plotted somewhat differently.

When an input signal is applied to the helix of a traveling-wave tube, all three forward waves are set up. The increasing wave grows until it predominates, and it forms the amplified output of the tube.

The total ac power of the increasing wave is zero. How can we obtain power from it? In the increasing wave we have a positive electromagnetic power flow in the circuit and an equal negative power flow in the electron stream. If we terminate the helix we can draw off the electromagnetic power; the electron stream is left with less power than it had on entering the helix.

DOUBLE-STREAM AMPLIFIERS

A double-stream amplifier makes use of two streams of electrons which have different velocities, as shown in Fig. 12. The behavior of a double-stream amplifier is very similar to that of a traveling-wave tube. In such a device each electron stream supports a slow, negative-energy wave and a fast, positive-energy wave. At a constant frequency ω let the velocity u_1 of one stream be kept constant and let the velocity u_2 of the other stream be varied. The behavior of the phase constants β of the waves is shown qualitatively in Fig. 13. β_{s1} and β_{f1} are the phase constants of the slow and fast waves of the constant-velocity stream, and β_{s2} and β_{f2} are the phase constants of the slow and fast waves of the stream whose velocity is changed. There are two ranges of velocity u_2 for which gain is obtained; for u_2 a little larger or a little smaller than u_1 .

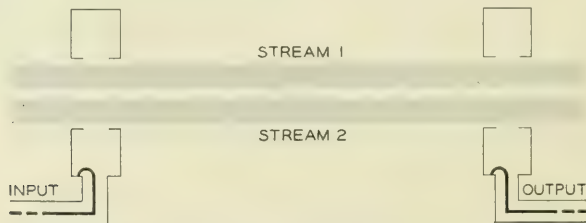


Fig. 12 — Two nearly electron streams of different velocities u_1 and u_2 constitute a double-stream amplifier.

NOISE WAVES ON ELECTRON STREAMS

Consider the electrons of the beam as they leave the cathode. If the velocity distribution is Maxwellian, and if the electrons leave independently, there will be a mean-square fluctuation in convection current, i^2 , given by

$$i^2 = 2eI_0B$$

and an uncorrelated mean square fluctuation in ac velocity, v^2 , given by

$$v^2 = (4 - \pi) \left(\frac{e}{I_0} \right) \left(\frac{kT_c}{m} \right) B$$

Here I_0 is beam current, e and m are electron charge and electron mass, k is Boltzmann's constant, T_c is cathode temperature and B is bandwidth.

Usually, space-charge-limited flow is used. In this case the beam current is only a part of the emitted current; the rest is turned back at the potential minimum. In this case we may use the above relations, counting I_0 as the beam current, as some sort of approximation for the current passing the potential minimum.

The wave picture we have been discussing may be seriously inaccurate near the cathode where the relative spread in electron velocities is large. Suppose that we hope for the best and apply it. We find that in the most general case our electron stream will have on it a noise standing-wave pattern. If $i_{\min}$ and $i_{\max}$ are the minimum and the maximum noise currents,

$$\frac{|i_{\min}| + |i_{\max}|}{2eI_0B} = \frac{1}{2} \alpha \left(\frac{\omega}{\omega_q} \right) \left(\frac{kT_c}{eV_0} \right)$$

Here α is a constant near to unity.

The noise pattern is made up of two uncorrelated noise standing-wave patterns, one from i at the cathode and the other from v at the cathode; these patterns have amplitudes i_1 and i_2 at their maxima; the minima are of course zero. We have

$$|i_{\min}| + |i_{\max}| = |i_1| + |i_2| \sin \Psi$$

Here Ψ is the relative phase angle of the standing-wave patterns associated with i_1 and i_2 . That is, if the maximum of the i_2 pattern is at that of i_1 , $\Psi = 0$, while if the maximum of the i_2 pattern is midway between maxima of i_1 , then $\Psi = \pi/2$.

The first of these theorems says something about the noise current at the maximum and that at the minimum, but it does not directly say how

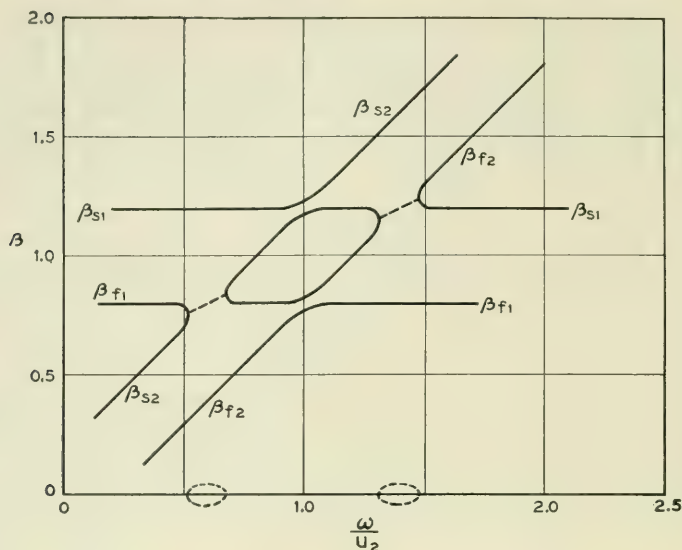


Fig. 13 — In a double-stream amplifier, gain is obtained when the phase constants of the slow wave of the faster stream and the fast wave of the slower stream are nearly equal.

large the maximum is. For an ordinary two-potential electron gun, $i_{\max}$ is very large compared with $i_{\min}$.

NOISE DEAMPLIFICATION

Early traveling-wave tubes made use of a two-potential electron gun spaced a critical distance from the circuit, as shown in Fig. 14. More recently it has been found possible to reduce the noise figure considerably by the use of space-charge-wave amplification, as discussed in the section on "The Space-Charge-Wave Amplifier." The structure used is indicated in Fig. 15. The gun has a low-potential anode followed by a

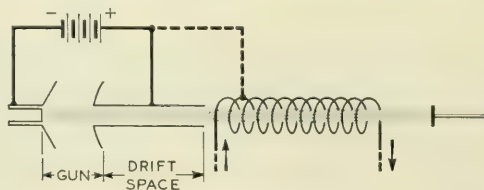


Fig. 14 — When a simple, two-potential electron gun is used, the noise figure of a traveling-wave tube can be optimized by adjusting the drift-space between the gun anode and the helix.

drift tube. At the point where the noise current is a minimum the voltage is "jumped" to the helix voltage. A second drift tube follows, so that there is a critical distance between the jump and the helix.

The effect of this "voltage jump" gun is to deamplify the component of the space-charge waves which is associated with the noise current at the current maximum. In space-charge-wave amplifier terms, this component sets up the decreasing wave only. Thus, in the second drift tube the ratio $|i_{\max}/i_{\min}|$ is smaller than in the first.

By using a single velocity jump, traveling-wave tubes with noise figures around 8 db have been made.

The use of more velocity jumps has been proposed. It can be shown, however, that as $i_{\max}$ is deamplified, $i_{\min}$ must be amplified. This sets a theoretical limit of around 6 db to the noise figure attainable by means of space-charge-wave deamplification alone.

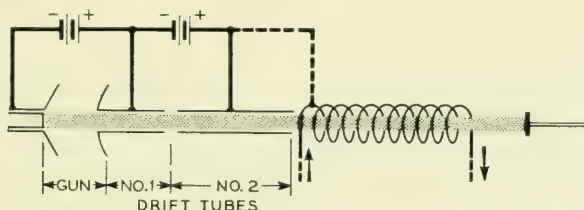


Fig. 15 — When a two-potential or "velocity jump" gun is used, the noise figure can be reduced by space-charge-wave deamplification of the noise on the electron stream.

NOISE CANCELLATION

It would be highly desirable to build a traveling-wave tube such that the electromagnetic input would excite an increasing wave, but the noise in the electron stream would excite only some combination of the decreasing and the unattenuated waves. If we succeeded in this, the noise introduced by the tube could be made as small relative to the signal as desired, merely by making the tube long enough. Can we accomplish this by means of some special structure near the input end of the tube?

We can represent the noise on the electron stream at some reference point by means of a velocity fluctuation v and a current fluctuation i ; we have seen that neither can be zero. Because the system is linear, superposition applies, and the amplitudes of the growing, attenuated and unattenuated waves which are excited are the sums of the amplitudes excited by i and v independently.

Suppose that $v = 0$. Then the beam carries no power. Thus, i cannot

excite the unattenuated wave, for that wave carries power. Let us assume that it excites the decreasing wave alone, which, when present alone, carries no power. So far there is no contradiction, and we can believe that it is possible to arrange matters so that the current alone does not excite the increasing wave.

Suppose we have so arranged matters that i does excite the decreasing wave only. Consider what happens when $i = 0$. Can v alone excite the decreasing wave only if i alone excites the decreasing wave only? If it can, then v and i together must excite the decreasing wave only. But suppose v and i are of the same frequency and in phase. Then the beam carries power. But, the decreasing wave alone cannot carry power, and hence what we have assumed is impossible. If the i excites the decreasing wave only, then v must excite at least a component of the growing wave. Hence, we cannot cancel out the noise from the beam completely.

FINAL COMMENTS

We have seen that the properties of space-charge waves and the behavior which must follow when space-charge waves are coupled to other space-charge waves or to circuit waves can be used to explain the operation of seemingly diverse types of microwave tubes. The wave picture gives a clear and quantitative picture of energy relations and power flow. It enables us to understand simply the effect of thermal velocities on the operation of tubes through their effect on the phase constants of the space-charge waves. It is useful in detailed considerations of noise, and in one case it has enabled us to draw a general conclusion without resorting to formal mathematical manipulation. It may well be that the wave picture can be of further use both in calculating detailed behavior of tubes and in understanding their general properties.

APPENDIX A

SPACE-CHARGE WAVES

Consider a narrow electron stream in which we may assume that electron velocity and charge density do not vary across the stream, and in which the electrons are free to move in the z -direction only. An axially symmetrical electron focusing system immersed, cathode and all, in a strong magnetic field approximates this.

Let all ac quantities contain the factor

$$e^{-j\beta z} e^{j\omega t}$$

and let the total charge density, current and electron velocity be made

up of dc and ac parts as follows:*

charge density: $-\rho_0 + \rho$

convection current density: $-I_0 + i$

velocity: $u_0 + v$

Here ρ_0 , I_0 and u_0 are positive dc quantities. The quantities on the right are the ac components.

We have from the definition of convection current

$$(-I_0 + i) = (-\rho_0 + \rho)(u_0 + v) \quad (\text{A1})$$

In the case of very low level operation, we neglect products of ac quantities in comparison with products of ac and dc quantities. Doing this, we obtain from (A1) the dc and ac convection currents

$$I_0 = \rho_0 u_0 \quad (\text{A2})$$

or

$$i = -\rho_0 v + u_0 \rho \quad (\text{A3})$$

$$\rho = \frac{i + \rho_0 v}{u_0} \quad (\text{A4})$$

We can apply the continuity equation, or, the equation of conservation of charge, to the ac convection current

$$\frac{\partial i}{\partial z} = -\frac{\partial \rho}{\partial t} \quad (\text{A5})$$

$$\beta i = \omega \rho$$

$$\beta i = \left(\frac{1}{u_0} \right) (j\omega i - \rho_0 j\omega v) \quad (\text{A6})$$

$$i = \frac{-\omega \rho_0 v}{\omega - \beta u_0}$$

Thus, if we have a wave with a given phase constant β , and if we know ρ_0 and ω , (A6) gives the convection current in terms of ac electron velocity. How can we find what β will be? To find this we must consider the effect of the electric field on the electrons. Consider an ac electric field E_z in the z direction, which also varies with time and distance as

* It will be convenient elsewhere to use $-I_0$ and i as currents rather than current densities and $-\rho_0$ and ρ as charge per unit length.

do the other ac quantities. We can write

$$\frac{dv}{dt} = -\frac{e}{m} E_z \quad (\text{A7})$$

Here e/m , the charge-to-mass ratio of the electron, is taken as a positive quantity.

In (A7), dv/dt is the rate of change of v with respect to t for a single electron; that is dv/dt observed riding along with the electron. If we ride along with the electron for a time dt we move along distance dz

$$dz = (u_0 + v) dt$$

For small signals we neglect v in this expression and write

$$dz = u_0 dt$$

Hence, the total change dv in the velocity of the electron in the time dt is

$$dv = \frac{\partial v}{\partial t} dt + \frac{\partial v}{\partial z} u_0 dt$$

Hence, we find that

$$\frac{dv}{dt} = j(\omega - \beta u_0)v \quad (\text{A8})$$

Using (A7) and (A8), we see that

$$v = \frac{j \frac{e}{m} E_z}{(\omega - \beta u_0)} \quad (\text{A9})$$

We can combine (A9) with (A6) and write for the convection current density

$$i = \frac{-j \frac{e}{m} \omega \rho_0 E_z}{(\omega - \beta u_0)^2} \quad (\text{A10})$$

Let us now consider a special, hypothetical case in which the electric field is in the z direction only, so that there are no transverse electric

fields and no transverse displacement current. Then the total ac current density i_t is the sum of the convection current density and the displacement current density, or,

$$i_t = i + j\omega\epsilon E_z$$

$$i_t = \left(\frac{-\frac{e}{m}\rho_0}{\epsilon(\omega - \beta u_0)^2} + 1 \right) j\omega\epsilon E_z \quad (\text{A11})$$

Let us use a quantity ω_p , which was long ago named the plasma frequency (radian frequency)

$$\omega_p^2 = \frac{\frac{e}{m}\rho_0}{\epsilon} \quad (\text{A12})$$

Using ω_p , (A10) can be written as

$$i_t = \left(-\frac{\omega_p^2}{(\omega - \beta u_0)^2} + 1 \right) j\omega\epsilon E_z \quad (\text{A13})$$

According to Maxwell's equations the divergence of the total current is zero. Both components of i_t vary with z . If, as we have assumed, there is no current normal to the z direction, then i_t must be zero. If this is to be so, we must have

$$(\omega - \beta u_0)^2 = \omega_p^2$$

$$\beta = \frac{\omega}{u_0} \pm \frac{\omega_p}{u_0} \quad (\text{A14})$$

In actual electron beams there is transverse electric field away from the beam, and hence i_t is not zero. It is found, however, that when ω_p is small compared with ω , we can write quite accurately

$$\beta = \frac{\omega}{u_0} \pm \frac{\omega_q}{u_0} \quad (\text{A15})$$

Here ω_q , which is known as the *effective plasma frequency*, is smaller than ω_p . As ω is raised, so that the wavelength of the space-charge waves becomes smaller compared with the diameter of the electron beam, the electric field tends to become largely longitudinal and ω_q approaches ω_p .

The upper sign in (A15) gives the phase constant of the slow wave, a wave with a phase velocity less than that of the electrons. The lower sign gives that of the fast wave, a wave with a phase velocity faster than that of the electrons.

From (A15) and (A6) we note that

$$i = \pm \frac{\omega}{\omega_q} \rho_0 v \quad (\text{A16})$$

The upper sign holds for the slow wave; the lower sign for the fast wave.

It has been convenient to use $-I_0$ and i as current densities and $-\rho_0$ and ρ as charge densities. In subsequent work and in the text, $-I_0$ and i will be used as beam current and $-\rho_0$ and ρ as charge per unit length. All the relations of this appendix except (A11)–(A13) will hold if the quantities are so interpreted.

APPENDIX B

POWER FLOW IN SPACE CHARGE WAVES

The purpose of this appendix is to justify the expression for power flow in the beam.

Consider that the electron beam is acted on over a short distance by an ac voltage. Imagine, for instance, that the beam passes through two very closely spaced grids which form a part of a resonator, and that a voltage ΔV appears between the grids. What does the voltage do to the beam?

The voltage ΔV changes the velocity of the electrons but it does not change the convection current. To find out how much the velocity is changed we need only consider the case in which the beam has no ac velocity on reaching the grids, since in a linear system the change will be the same in all other cases. The total velocity $u_0 + v$ is given in terms of the total accelerating voltage $V + \Delta V$ by

$$u_0 + v = \sqrt{2 \frac{e}{m} (V_0 + \Delta V)} \quad (\text{B1})$$

We assume ΔV to be small, so that

$$v = \frac{\Delta V \frac{e}{m}}{\sqrt{2 \frac{e}{m} V_0}} = \frac{\Delta V \frac{e}{m}}{u_0} \quad (\text{B2})$$

The change ΔU in the "voltage" U is

$$\Delta U = - \frac{u_0 v}{\frac{e}{m}} = -\Delta V \quad (\text{B3})$$

The convection current i flows against the voltage ΔV , so that a power ΔP is transferred from the beam to the resonator which is attached to the grids.

$$\Delta P = -Re\Delta V i^* \quad (\text{B4})$$

Thus, the change in the power in the beam in passing through the grids must be $-\Delta P$

$$-\Delta P = Re(-\Delta V i^*) = Re\Delta U i^* \quad (\text{B5})$$

$$\Delta P = -Re\Delta U i^*$$

According to the expression we have used in calculating beam power, if the "voltage" of the beam on reaching the grids is U , and the convection current is i , then the beam power P_1 on reaching the grids is

$$P_1 = ReU i^* \quad (\text{B6})$$

After passing through the grids, U is increased by an amount ΔU while the current is unchanged, so that the power P_2 of the beam leaving the grids is

$$P_2 = Re(U + \Delta U) i^* \quad (\text{B7})$$

The loss of power in the beam, ΔP , is

$$\Delta P = P_1 - P_2 = -Re\Delta U i^* \quad (\text{B8})$$

This agrees with (B5), in which ΔP was calculated as the power lost from the beam to the resonator.

APPENDIX C

THE EFFECT OF THE VELOCITY DISTRIBUTION IN THE ELECTRON BEAM ON THE EFFECTIVE PLASMA FREQUENCY

Consider an electron beam in which electron motion is confined to the z -direction, and in which the electrons have a velocity spread with a

mean square deviation $\langle u^2 \rangle$ about the mean value u_0 . If ω_{q0} is the effective plasma frequency for $\langle u^2 \rangle = 0$, then taking the velocity spread into account, the effective plasma frequency ω_q is given approximately by

$$\omega_q^2 = \omega_{q0}^2 \left(1 + 3 \frac{\langle u^2 \rangle}{u_0^2} \left(\frac{\omega}{\omega_{q0}} \right)^2 \right)$$

If the velocity distribution is the same as for electrons accelerated individually by a voltage V_0 , that is, if electron interactions do not affect the velocity distribution appreciably (as they probably do not)

$$\frac{\langle u^2 \rangle}{u_0^2} = \frac{1}{4} \left(\frac{kT_c}{eV_0} \right)^2 = \frac{1}{4} \left(\frac{T_c}{11,600 V_0} \right)^2$$

Here k is Boltzman's constant and T_c is cathode temperature. Thus, from this assumption

$$\omega_q^2 = \omega_{q0}^2 \left(1 + \frac{3}{4} \left(\frac{kT_c}{eV_0} \right)^2 \left(\frac{\omega}{\omega_{q0}} \right)^2 \right)$$

Following our wave picture, we can take into account the thermal velocity spread by using this corrected value for the effective plasma frequency in all our formulae. For all practical purposes, the change in effective plasma frequency due to thermal velocities is negligible.

In a paper which will appear in the Journal of Applied Physics, D. A. Watkins has used a somewhat different approach in treating the effect of thermal velocities on the operation of traveling-wave tubes.

APPENDIX D

PHASE AND ATTENUATION CURVES FOR COUPLED MODES

When two unattenuated modes of propagation are coupled together periodically in a lossless manner, they combine to form two new modes. For each of these new modes the amplitude is changed in one period of the coupling structure by a factor

$$Me^{-j(\theta_q - \theta_p)/2} e^{j(\theta_3 + \theta_1)/2} \quad (D1)$$

where M is a root of

$$M^2 - 2\sqrt{1 \mp k^2} \cos \frac{(\theta_q - \theta_p + \theta_1 - \theta_3)}{2} + 1 = 0 \quad (D2)$$

Here k is a coupling coefficient which is zero for zero coupling. The upper sign applies if the power flow in the two modes have the same

signs while the lower sign applies if the power flows have opposite signs. θ_q and θ_p are phase lags per coupling period associated with the two original modes and θ_1 and θ_2 are phase angles associated with the coupling device.

We can treat the case of continuous coupling by letting the period of coupling L be very short, the angles θ_q , θ_p , θ_1 , θ_3 be very small, and the coupling per period, k , be very small. In this case the cosine can be represented by the first terms of a power series and we find that the phase constants β of the modes are given by

$$\beta = \frac{\beta_a + \beta_b}{2} \pm \left(\frac{\beta_a - \beta_b}{2} \right) \sqrt{1 \pm \left(\frac{2K}{\beta_a - \beta_b} \right)^2} \quad (\text{D3})$$

Here β_a and β_b are the phase constants for $K = 0$ (zero coupling)

$$\beta_a = \frac{\theta_p - \theta_1}{L} \quad (\text{D4})$$

$$\beta_b = \frac{\theta_q - \theta_3}{L} \quad (\text{D5})$$

and K is the coupling per unit length

$$K = \frac{k}{L} \quad (\text{D6})$$

As before, the upper sign in the radical applies when the power flows have the same signs and the lower sign when the power flows have opposite signs.

In applying (D3) to the case of traveling-wave tubes and backward-wave oscillators, the effect of all but two modes was of course neglected when the two phase constants would have had nearly the same value in the absence of coupling; the curves for such regions were then joined smoothly to give the overall plots of Figs. 11 and 13.

In Fig. 11 the parameters chosen arbitrarily were:

$$\beta_c = 1$$

$$\beta_s = \omega/u_0 + 1/2$$

$$\beta_f = \omega/u_0 - 1/2$$

$$K = 0.1$$

The complex portion of the phase constant, or, the real portion of the propagation constant, in a stop band caused by the coupling of two

modes with power flows of opposite signs is designated by α and is plotted as the dashed ellipses about the horizontal axis.

APPENDIX E

This appendix comments briefly on various sections and cites references, which are listed at the end of the appendix. The list of references is not exhaustive, but it should enable the interested reader to follow work back to its source.

1. *Space-Charge Waves*

Space-charge waves of the general sort considered are related to the plasma oscillations of Tonks and Langmuir.¹ Waves in long beams were first discussed by Hahn² and Ramo.³ The effects of a velocity distribution are discussed by Pierce⁴ and by Bohm and Gross.⁵ The negative energy of the slow space-charge wave has been reported by Chu⁶ and by Walker.⁷ Chu gave the effective "voltage" U and the characteristic impedance K for the waves.

2. *The Klystron*

Beck⁸ gives an adequate description of and references to klystrons.

3. *The Resistive Wall Amplifier*

The effect has been discussed by Pierce,⁹ and a tube using it has been described by Birdsall, Brewer and Haeff.¹⁰

4. *The Easitron; Increasing Wave in a Lossless System*

The original easitron was a tube built by L. R. Walker at Bell Telephone Laboratories; it was a 3-cm tube using half-wave wires as resonant elements. It has not been described in the literature. Pierce has discussed the operation of this sort of multi-resonator klystron on page 195 of *Traveling Wave Tubes*¹¹ and elsewhere.⁹

5. *Coupling of Modes of Propagation*

The operation of traveling-wave tubes was first explained in terms of coupling between an electromagnetic wave and a space-charge wave by C. C. Cutler in unpublished work. Mathews has made an analysis in

these terms.¹³ Such coupling has been considered in general terms by Pierce.¹²

6. *The Space-Charge-Wave Amplifier*

This tube was invented by Tien, Field and Watkins¹⁴ and is described in more detail by Tien and Field.¹⁵

7. *The Traveling Wave Tube*

Adequate descriptions and references are available in work by Kompfner,¹⁶ Pierce,¹¹ and Beck.⁸

8. *Double-Stream Amplifiers*

Descriptions and references are given by Pierce¹¹ and by Beck.⁸

9. *Noise Waves in Electron Streams*

Cutler and Quate have published experimental results.¹⁷ The theorems quoted are given by Pierce.¹⁸

10. *Noise Deamplification*

This was suggested by Tien, Field and Watkins¹⁴ and is described in detail by Watkins¹⁹ and Peter.²⁰

11. *Noise Cancellation*

Noise cancellation was first proposed by C. F. Quate.²¹

REFERENCES

1. Lewi Tonks and Irving Langmuir, *Phys. Rev.*, **33**, pp. 195-210 and p. 990, 1929.
2. W. C. Hahn, *Gen. Elect. Rev.*, **42**, pp. 258-270, 1939.
3. Simon Ramo, *Phys. Rev.*, **56**, pp. 276-283, 1939.
4. J. R. Pierce, *Jour. App. Phys.*, **19**, pp. 231-236, 1948.
5. D. Bohm and E. P. Gross, *Phys. Rev.*, **75**, pp. 1851-1876, 1949, **79**, pp. 992-1001, 1950.
6. L. J. Chu, paper presented at the Institute of Radio Engineers Electron Devices Conference, University of New Hampshire, June, 1951.
7. L. R. Walker, *J. App. Phys.*, **25**, pp. 615-618, May, 1954.
8. *Thermionic Valves*, A. H. W. Beck, Cambridge University Press, 1953.
9. J. R. Pierce, *B.S.T.J.*, **30**, pp. 626-651, 1951.
10. Charles K. Birdsall, George R. Brewer and Andrew V. Haeff, *Proc. I. R. E.*, pp. 865-871, 1953.
11. *Traveling Wave Tubes*, J. R. Pierce, Van Nostrand (1950).
12. W. E. Mathews, *J. App. Phys.*, **22**, pp. 310-316, 1951.

13. J. R. Pierce, *J. App. Phys.*, **25**, pp. 179-183, Feb. 1954.
14. Ping King Tien, Lester M. Field and D. A. Watkins, *Proc. I.R.E.*, **39**, p. 194, 1951.
15. Ping King Tien and Lester M. Field, *Proc. I.R.E.*, **40**, pp. 688-695, 1952.
16. R. Kompfner, *Rep. Progress. Phys.*, **15**, pp. 275-327, 1952.
17. C. C. Cutler and C. F. Quate, *Phys. Rev.*, **80**, pp. 875-878, 1950.
18. J. R. Pierce, *J. App. Phys.*, **8**, pp. 93-933, 1954.
19. D. A. Watkins, *Proc. I.R.E.*, **40**, pp. 65-70, 1952.
20. R. W. Peter, *R.C.A. Review*, **13**, pp. 344-368, 1952.
21. C. F. Quate, paper presented at the Institute of Radio Engineers Electron Devices Conference, University of New Hampshire, 1952.

Theory of Open-Contact Performance of Twin Contacts

By M. M. ATALLA and MISS R. E. COX

(Manuscript Received June 17, 1954)

The first part is a presentation of an analytical study of the open-contact performance of twin contacts. It provides means for predicting their performance from single contact data. It is shown that the probability of failure of twin contacts is generally appreciably greater than the square of the probability of failure of single contacts. This is supplemented with the results of an experimental study which determines the effects of a few design parameters on the performance of single contacts. These are the parameters that determine the magnitude of improvement in performance obtained by replacing single contacts by twin contacts.

INTRODUCTION

The present switching apparatus normally operates in atmospheres that may be contaminated with dust particles and foreign matter. Some apparatus components, particularly the contacts, are relatively sensitive to such contaminations which may interrupt the proper functioning of a pair of contacts. Normally, a single switching operation in a central office requires the operation of as many as a thousand relays or 10,000 contacts. To secure the high level of performance desired, it is evident that superlative performance and high degree of reliability of the contacts are essential.

Many attempts have been and are being made to reduce the so-called "open" contact troubles due to foreign matter. Examples of environmental precautions are filtering the air supply to the central office, enclosing apparatus in cabinets, limiting personnel activities in the office, etc. An additional precaution incorporated in the apparatus design is the use of twin contacts. Such a scheme, when *properly* used, should result in substantial improvement in performance since an open can only take place when both members become open simultaneously. It may occur to one that the probability of a twin-contact open is the square of the prob-

ability of a single-contact open. Such a performance, however, has never been observed in practice where an improvement of only 10:1 is usually more typical. A study of the mechanisms involved has revealed that only a very small number of opens are obtained due to the simultaneous occurrence of opens on both members of a twin contact. The majority of opens, however, occur by having an open in one member of a twin contact which persists long enough to allow the occurrence of an open on the other member. By expressing this physical process in mathematical terms it was possible to develop a theory of performance of twin contacts in terms of the characteristics of single contacts.

NOTATION

d	Diameter of dust particle
f	Fractions of opens in single contacts cleared after N operations
f_{∞}	The asymptotic value of f corresponding to $N = \infty$
$\bar{n}$	Average number of operations required to clear an open on a single contact
r	Distance of particle from center of circular open contact zone
r_0	Radius of "open zone"
s	Fraction of the twin contacts that are half open at any time, $s = s_{\infty} + s_n$
s_{∞}	Fraction of twin contacts that are permanently half open
s_n	Fraction of twin contacts that are temporarily half open
w	Mechanical wipe
x	Average displacement of a dust particle per contact operation
F	Contact force
N	Number of contact operations
P_s	Probability of occurrence of a single-contact open in opens/contact operation
P_t	Probability of occurrence of a twin-contact open in opens/contact operation
X	Total displacement distance to clear an open
α	$= (2 - P_s)(1 - f_{\infty})$
β	$= \bar{n}f_{\infty}P_s(2 - P_s)$
θ	Angle of displacement
φ	Slope of contact surface irregularity

PRESENTATION OF THEORY

Outline and Assumptions

Consider a large group of twin contacts each constituting a pair of identical and entirely independent single contacts. After a period of

operation a certain number of the twin contacts will become half-open.* These contacts will behave as if they were single-contacts until either: (1) the open half clears itself by operation, or (2) the other half becomes open leading to a twin-contact open. It is assumed that a failing twin-contact is cleared by an operator and then put back into service.

In developing the theory it was necessary to represent by analytic expressions the rate of occurrence of opens on single contacts and the rate of their clearing by operation. These are approximations of a fairly large amount of experimental data consistently obtained from a number of tests on a variety of actual telephone relay contacts.

(1) Rate of opens of single contacts " P_s ": For a large number of single contacts at one set of operating conditions, the rate of opens is usually constant. This constant depends primarily on the quality and concentration of the offending foreign matter involved, the design of the contacts and their mechanism of actuation. Fig. 1 shows the results of three tests on single contacts of different design at different test conditions. They all substantiate the assumption that the rate of opens of

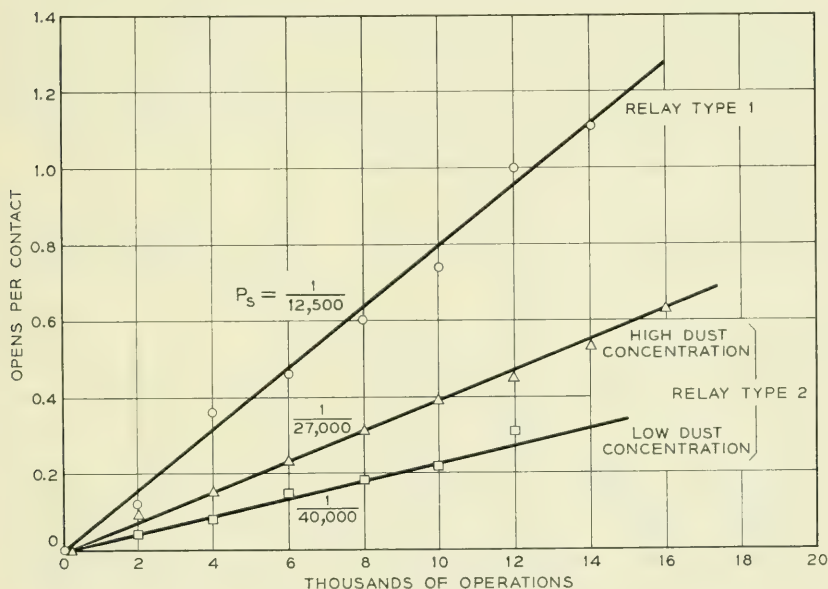


Fig. 1 — Rate of opens of single relay contacts.

* A half-open is defined as one where only *one* member, of a pair in a twin-contact, is open. In practice, a half-open is not normally detected. Only a simultaneous open on both members of a twin contact will cause a circuit failure.

single contacts is a constant:

$$P_s = \text{constant} \quad (1)$$

where P_s is defined as the number of opens per "contact operation." In some cases, a certain transient period may precede the equilibrium characteristic " $P_s = \text{constant}$ characteristic." This transient period is usually relatively short and is neglected in this analysis.

(2) Clearing of opens on single-contacts by operation: If an open single contact is allowed to operate mechanically, it is possible that it will clear itself after a number of operations. As discussed in a later section, the opens obtained are never identical in nature. They, instead, have a certain statistical distribution which usually accounts for a wide spread in their clearing rate. If, however, the operating conditions are under control, the clearing characteristic of a set of contacts is found to follow a well defined and reproducible statistical distribution. Fig. 2 shows an accumulative distribution curve for clearing opens produced by cotton lint fibres.* The ordinate represents the fraction of the open single contacts that clear after N operations as given by the abscissa. In general, these relations have the following typical characteristics. The first operation following the occurrence of the open is the most efficient† single operation in clearing opens. It is usually responsible for clearing 10 to 30 per cent of the total number of opens. The subsequent operations are progressively less efficient and in general a certain fraction

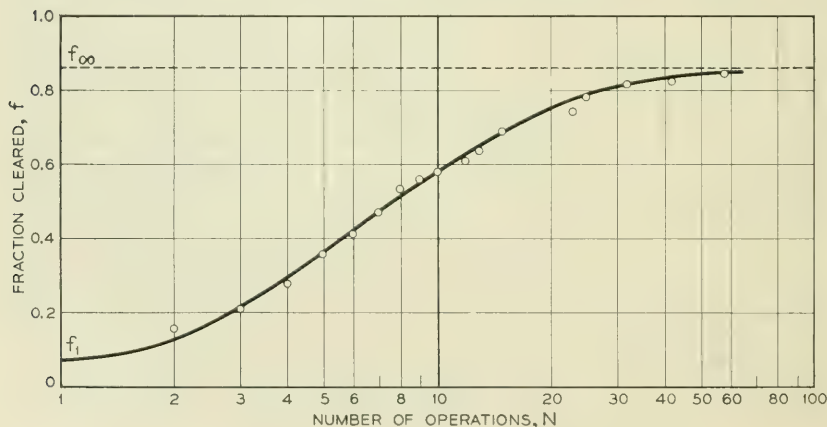


Fig. 2 — Distribution of clearing opens caused by lint.

* This is one of the major causes of open contacts in central offices. These fibres are usually in ribbon form of various configurations.

† This apparent efficiency is only due to the presence of opens that are more easy to clear than others. These will readily clear after one or a few operations.

$(1 - f_\infty)$ of the opens will persist for a relatively large number of operations. A study of a variety of these clearing characteristics generally indicates a rapid rise to the asymptotic value f_∞ in less than 100 operations, and to $f = 0.5$ in less than 10 operations. As will be shown the fractional persistency $(1 - f_\infty)$ is of major importance in determining twin-contact performance. In general, however, opens on twin-contacts are due to both the persistent half-opens and the temporary half-opens that might develop into twin-opens before clearing takes place.

DEVELOPMENT OF THE THEORY

Consider a large number of contacts operating at steady conditions. After N operations, let s_∞ be the fraction of the contacts that is permanently half-open and $s_{\bar{n}}$ be the fraction that is temporarily half-open. As discussed, the number of operations necessary to clear a half-open is not constant and, for the majority of the contacts, is of the order of a few operations. To simplify the treatment, it is assumed that each temporary half-open will clear in an average of $\bar{n}$ operations *from the time it first occurred*. The fraction $s_{\bar{n}}$ must, therefore, have been produced during the $\bar{n}$ operations directly preceding the time t . Since $\bar{n}$ is usually relatively small, the universe can be assumed to have had a negligible change during the operations $\bar{n}$. Hence,

$$\begin{aligned}s_{\bar{n}} &= (\text{rate of formation of temporary half-opens}) \times \bar{n} \\ &= \bar{n}[2(1 - s)P_s - (1 - s)P_s^2]f_\infty\end{aligned}$$

where s = total fraction of half-opens = $s_{\bar{n}} + s_\infty$. Substituting $\beta = \bar{n}f_\infty P_s(2 - P_s)$ one gets

$$s_{\bar{n}} = \frac{\beta}{1 + \beta} (1 - s_\infty) \quad (2)$$

Also, after N operations, the incremental change ds_∞ due to dN operation is:

$$ds_\infty = [2(1 - s)P_s - (1 - s)P_s^2](1 - f_\infty) dN - s_\infty P_s dN$$

where the second term is the reduction in s_∞ due to occurrence of twin-opens. Substituting $\alpha = (2 - P_s)(1 - f_\infty)$ and combining with equation 2 to eliminate $s_{\bar{n}}$ give:

$$ds_\infty = -\frac{1 + \alpha + \beta}{1 + \beta} \left[s_\infty - \frac{\alpha}{1 + \alpha + \beta} \right] P_s dN \quad (3)$$

$$s_\infty = \frac{1}{1 + \alpha + \beta} + K e^{-(1+\alpha+\beta)/(1+\beta)P_s N} \quad (4)$$

where K is an integration constant. Let at $N = 0$, $s_\infty = s_0$, which is a description for the initial conditions of the contacts. The solution becomes:

$$s_\infty = \frac{\alpha}{1 + \alpha + \beta} \left[1 - \left(1 - \frac{1 + \alpha + \beta}{\alpha} s_0 \right) e^{-(1 + \alpha + \beta)/(1 + \beta) P_s N} \right] \quad (5)$$

The rate, or probability, of twin-contact failure is determined from:

$$P_t = s P_s + (1 - s) P_s^2 \quad (6)$$

Substituting from 2 and 5 into 6 and reducing gives:

$$P_t = P_s^2 + P_s(1 - P_s) \frac{\beta}{1 + \beta} \left[1 + \frac{\alpha' \beta}{1 + \alpha + \beta} \left(1 - \left(1 - \frac{1 + \alpha + \beta}{\alpha} s_0 \right) e^{-(1 + \alpha + \beta)/(1 + \beta) P_s N} \right) \right] \quad (7)$$

where, one may repeat for convenience:

$$\alpha = (2 - P_s)(1 - f_\infty) \text{ and } \beta = \bar{n} f_\infty P_s (2 - P_s)$$

For all practical cases, $P_s \ll 1$, $\alpha = 2(1 - f_\infty)$ and $\beta = 2\bar{n} f_\infty P_s < 1$. Substituting in 7 gives

$$P_t = P_s^2 + 2P_s \left[\bar{n} f_\infty P_s + \frac{1 - f_\infty}{3 - 2f_\infty} \left(1 - \left(1 - \frac{3 - 2f_\infty}{2(1 - f_\infty)} s_0 \right) e^{-(3 - 2f_\infty) P_s N} \right) \right] \quad (7')$$

This is a general expression, relating the expected performance of twin-contacts to that of single contacts. It is evident that the idealistic performance of $P_t = P_s^2$, i.e., the probability of a twin-contact failure is the square of that for single contacts, can only be achieved if: (a) $f_\infty = 1.0$, i.e., persistent half-opens never occur, (b) $\bar{n} = 0$, i.e., each temporary half-open occurring during one operation will clear during the subsequent operation, and (c) $s_0 = 0$, i.e., there is no initial contamination. These conditions are never obtained in practice and generally P_t is much greater than P_s^2 .

Equation 7' also indicates that at the beginning of operation, when the exponent is much less than 1.0, P_t is given by:

$$(P_t)_0 = P_s^2(1 + 2\bar{n} f_\infty) + P_s s_0 \quad (8)$$

Numerically if $\bar{n} = 50$, $f_\infty = 1.0$ and $s_0 = 0$, $P_t = 101 P_s^2$ which is 101 times worse than the idealistic performance of $P_t = P_s^2$. The initial rate

of failure of twin contacts is also quite sensitive to initial contact contamination. If, for example, $s_0 = 10^{-3}$, i.e., $1/1000$ of the twin contacts are permanently half-open to start with, and for the same numbers used above and $P_s = 1/10^6$, equation 8 gives $P_t = 1.1 \times 10^{-9}$. This corresponds to an 11 fold increase in twin-contact failures just due to an initial contamination s_0 of 0.1 per cent. This performance is also 1100 times worse than the idealistic performance of $P_t = P_s^2 = 10^{-12}$.

By operation, the performance of twin contacts will exponentially deteriorate according to equation 7. It will asymptotically approach a constant rate of failure given by:

$$(P_t)_\infty = P_s^2(1 + 2\bar{n}f_\infty) + 2P_s \frac{1 - f_\infty}{3 - 2f_\infty} \quad (9)$$

This is independent of the initial contamination s_0 and is practically reached in a number of operations:

$$(n_t)_\infty = 3/(P_s(3 - 2f_\infty)) \quad (10)$$

The worst performance of twin contacts is obtained when $f_\infty = 0$, i.e.,

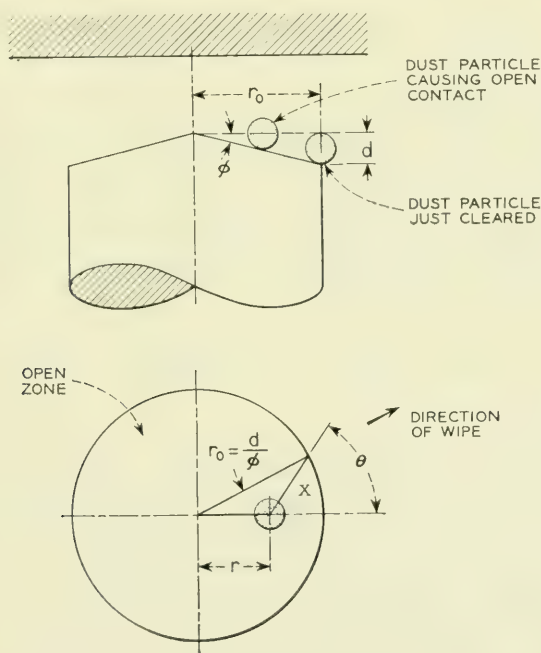


Fig. 3 — Particle in open zone.

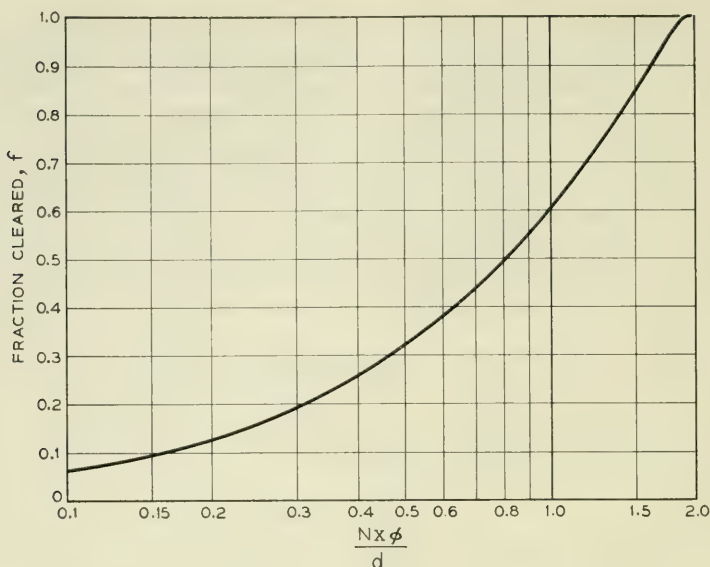


Fig. 4 — Accumulative probability distribution of fraction cleared “f” versus $(Nx\phi/d)$.

every half-open persists indefinitely, and is given by $(P_t)_\infty = \frac{2}{3}P_s$. In other words, by replacing single contacts by twin contacts the frequency of failures is decreased by only one third. In practice, however, f_∞ is rarely that low and under the wide variety of central office conditions it may range between 0.85 and 0.98. This corresponds to a range of performance of $P_t = 0.23 P_s$ to $0.038 P_s$ or the frequency of twin-contact failures is about $\frac{1}{4}$ to $\frac{1}{26}$ that of single contacts.

CLEARING AN OPEN BY MECHANICAL OPERATION

Introduction

Fig. 3 shows a diagrammatic sketch of the contact area with spherical particles preventing metallic contact. The surface irregularity has a slope ϕ and the particle diameter is d . It is evident that the particle will produce an open if it falls within a limiting radius $r_0 = d/\phi$. The area within r_0 is called the “open zone.” For a particle at a radius r within the open zone, mechanical wipe w will tend to displace the particle at an angle θ , $0 \leq \theta \leq 2\pi$. The open is cleared when the particle is displaced a distance X sufficient to drive it out of the open zone. For unidirectional

wipe, if the average displacement per operation is x , the number of operations required to clear an open is $N = X/x$. Since, however, the initial location of the particle r has a triangular probability distribution, $0 \leq r \leq r_0$, and the direction of wipe θ has a rectangular distribution, $0 \leq \theta \leq 2\pi$, the number of operations for clearing must have a corresponding probability distribution. This has been determined graphically and is shown in Fig. 4. This is an accumulative probability distribution of the fraction cleared " f " versus $(Nx\varphi/d)$. It indicates that 50 per cent of the opens will clear at $(Nx\varphi/d) = 0.8$ and 100 per cent at 2.0. The corresponding number of operations N can be determined only if x is known* under the operating conditions. By increasing the contact wipe w one may expect x , the average displacement per operation, to increase. Also by increasing the force, the frictional driving force will increase and x should also increase. One may, therefore, tentatively assume that x is proportional to $w^a F^b$ and the distribution function may be put in the form $f = f(Nw^a F^b)$, keeping all the other parameters fixed. This suggests that by varying the contact wipe w and the contact force F , the distribu-

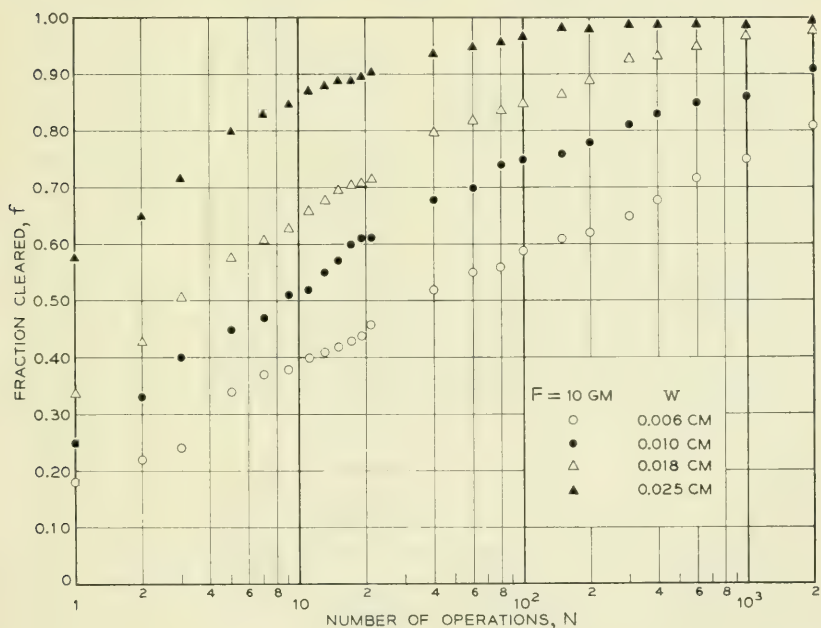


Fig. 5 — Effect of wipe.

* For a certain surface roughness φ and particle size d .

tion function f should be unique if plotted against Nw^aF^b where the a and b are constants to be determined experimentally.

Experiment

Cotton lint fibres and clean contact surfaces were used exclusively in this study. The fibres were essentially in the form of ribbons a few microns thick and of variable width and length. By using a dust separator,* lint fibres with a well controlled size distribution were collected on a glass plate. The setting used gave a rather uniform monolayer of fibres 80 per cent of which had a width between 10 and 20 microns.

The contacts tested were flat and made of palladium.† They were cleaned with methyl alcohol and distilled water, then dried. The collected lint fibres were transferred to the surface of one contact by a special adapter which allows the pressing of the contact on the glass

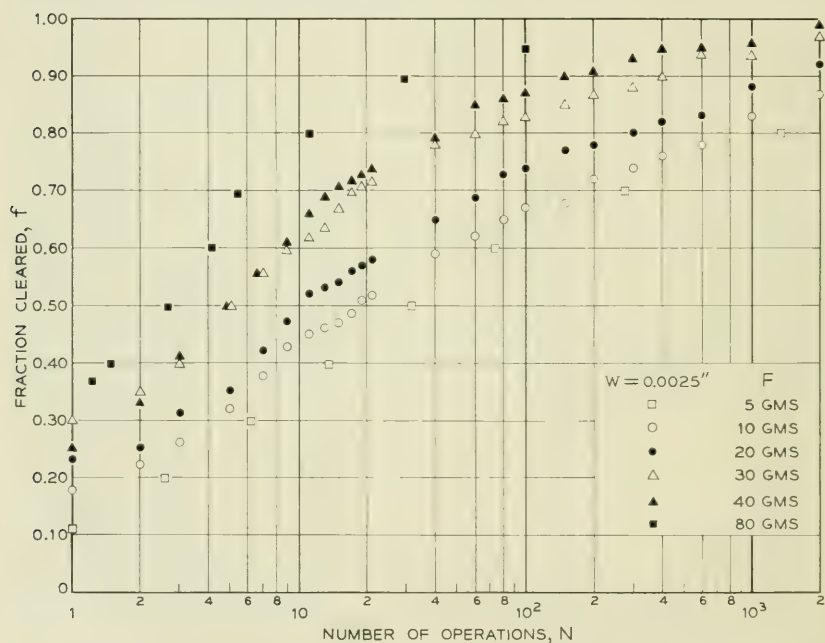


Fig. 6 — Effect of force.

* Based on controlling sedimentation by adjusting air speed in a two-stage separator.

† Contact surface roughness was controlled by frequently polishing the contact surfaces by a fixed process.

plate with the fibres. The pressing force was the same as that used in the subsequent operation of the contacts. The contacts were then operated at four operations per second in a sealed compartment. After each closure a checking circuit using 48 volts, and a maximum current of 0.50 amp., checked the continuity in the contacts. When the open was cleared the unit automatically stopped and the corresponding number of operations was obtained from a counter. The maximum number of operations allowed for each run was 2,000. For one set of operating conditions, it was necessary to repeat the above for at least 150 times before a representative clearing distribution was obtained.

Results

Effect of wipe: Fig. 5 shows the results obtained for a range of wipes between 0.006 and 0.025 cm at a constant force of 10 grams. As expected, the clearing rate was higher for larger wipes.

Effect of force: Fig. 6 was obtained at a constant wipe of 0.006 cm and a set of forces between 5 and 80 grams. Large forces gave higher clearing rates. The effect of changing the force, however, is not as significant as that of changing the wipe.

As outlined in the preceding introduction, the above data was replotted as fraction clearing f versus Nw^aF^b . The results are shown in Fig. 7 with $a = 3$ and $b = 1.0$. As indicated, the points converged to a single average line with comparatively small spread.* This shows that, at least for the range covered, the change in clearing rate obtained by changing the wipe say by a factor of two can also be obtained by changing the force by a factor of 8.

To determine the persistency $(1 - f_\infty)$, one may choose an arbitrary number of operations for defining it. If 2,000 operations is chosen, one may determine from the above data the fraction, $(1 - f_{2,000})$, that will persist to beyond 2,000 operations. This was done and the results are plotted in Fig. 8 as $(1 - f_{2,000})$ versus $F^{1/3}w$. This suggests the following relation:

$$(1 - f_{2,000}) = e^{-F^{1/3}w/0.006} \quad (11)$$

where F is in grams and w in cms. This expression allows the determination of the effects of force and wipe on the performance of twin contacts by substituting in Equations 7' through 10. Similarly the average number of operations $\bar{n}$, used in the above equations, may be obtained. This may

* This same convergence was obtained, but not presented here, by plotting f versus NF at constant w and f versus Nw^3 at constant F .

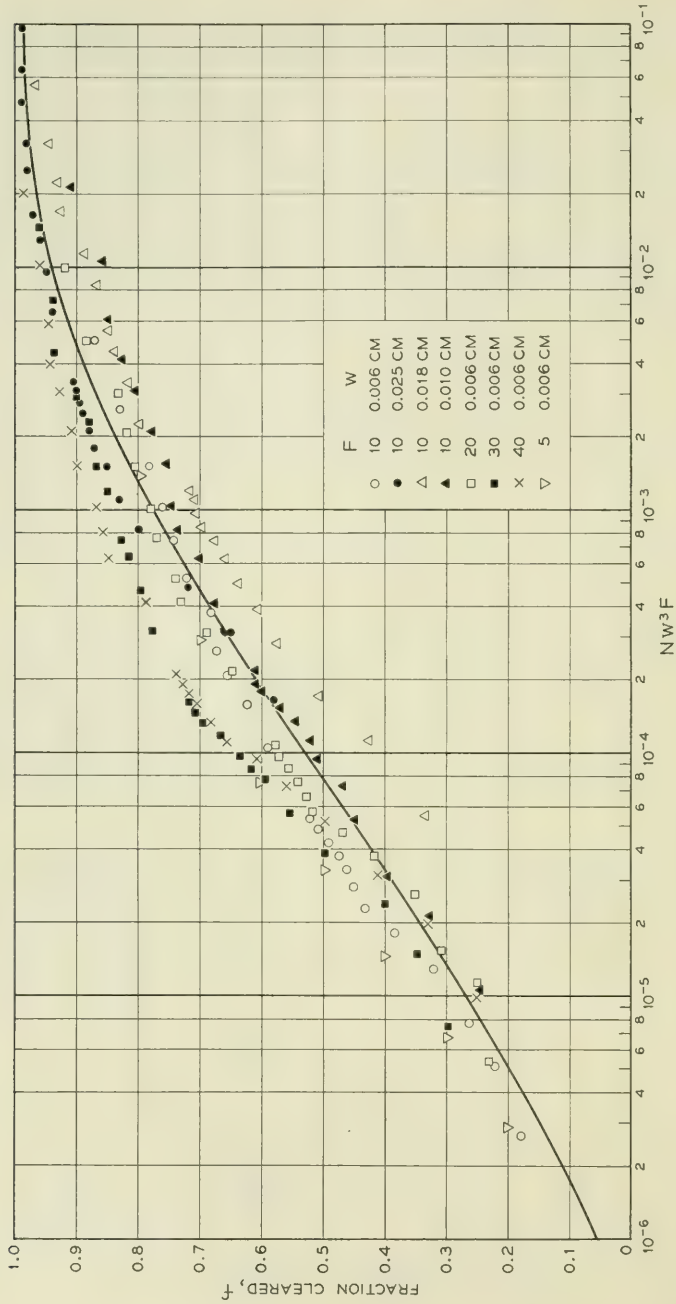


Fig. 7 — Average experimental curve; fraction cleared " f ," versus Nw^3F .

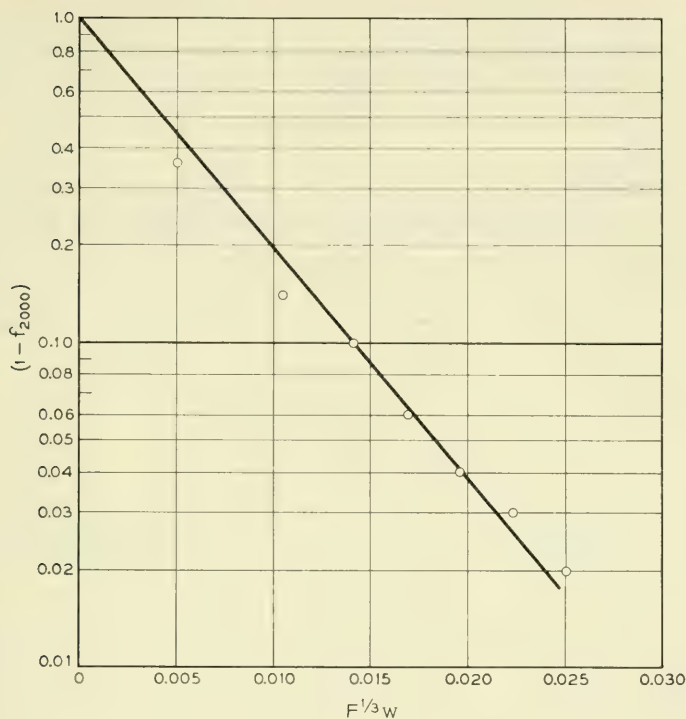


Fig. 8 — Persistency of opens at 2,000 operations.

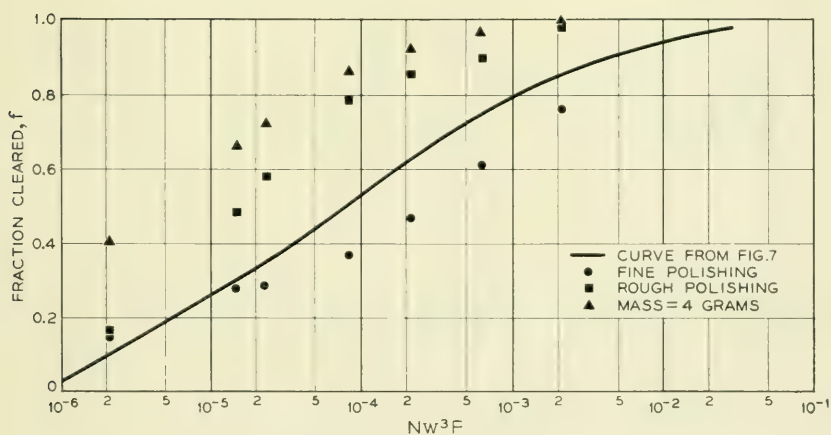


Fig. 9 — Effect of other parameters.

be arbitrarily defined as the number of operations at which 50 per cent of the opens will have cleared. At this or any other value of f one obtains from Fig. 7 $\bar{n}Fw^3 = \text{constant}$ or $\bar{n}$ is inversely proportional to Fw^3 .

OTHER EFFECTS

Fig. 9 shows the results obtained by varying other parameters. The solid line, obtained from Fig. 7 is shown for comparison. Indicated are the effects of fine polishing and rough polishing of the contact surfaces and of increasing the mass of the moving contact from 0.5 to 4 grams.

Bell System Technical Papers Not Published in this Journal

AHEARN, A. J., see HANNAY, N. B.

BEACH, A. L., see GULDNER, W. G.

BIDDULPH, R.¹

Short Term Autocorrelation Analysis and Correlatogram of Spoken Digits, J. Acous. Soc. Am., **26**, pp. 539-541, July, 1954.

BRATTAIN, W. H., see GARRETT, C. G. B.

BULLINGTON, K.¹

Reflection Coefficients of Irregular Terrain, Proc. I.R.E., **42**, pp. 1258-1262, Aug., 1954.

DEWALD, J. F.,¹ and LEPOUTRE, GERARD²

The Thermoelectric Properties of Metal — Ammonia Sodium and Potassium at -33° , J. Am. Chem. Soc., **76**, pp. 3369-3373, July, 1954.

FINE, M. E.,¹ and KENNEY, NANCY T.¹

Moduli and Internal Friction of Magnetite as Affected by the Low-Temperature Transformation, Phys. Rev., **94**, pp. 1573-1576, June 15, 1954.

¹ Bell Telephone Laboratories, Inc.

² Faculté Libre des Sciences, Lille, Nord, France.

FLETCHER, R. C.,¹ YAGER, W. A.,¹ PEARSON, G. L.,¹ and MERRITT, F. R.¹

Hyperfine Splitting in Spin Resonance of Group V Donors in Silicon, Letter to the Editor, *Phys. Rev.*, **95**, pp. 844-5, Aug. 1, 1954.

GAMBRELL, J. B., JR.¹

What is a Printed Publication Within the Meaning of the Patent Act, *J. Patent Office Society*, **36**, pp. 391-405, June, 1954.

GARRETT, C. G. B.,¹ and BRATTAIN, W. H.¹

Self Powered Semiconductor Amplifier, Letter to the Editor, *Phys. Rev.*, **95**, pp. 1091-1092, Aug. 15, 1954.

GREEN, E. I.¹

Creative Thinking in Scientific Work, *Elec. Eng.*, **73**, pp. 489-494, June, 1954.

GREMLING, R. C., see WRIGHT, MARIE G.

GULDNER, W. G.,¹ and BEACH, A. L.¹

Gasometric Method for Determination of Hydrogen in Carbon, *Analytical Chemistry*, **26**, pp. 1199-1202, July, 1954.

HANNAY, N. B.¹

A Mass Spectrograph for the Analysis of Solids, *Rev. Scient. Instr.*, **25**, pp. 644-648, July, 1954.

HANNAY, N. B.,¹ and AHEARN, A. J.¹

Mass Spectrographic Analysis of Solids, *Analytical Chemistry*, **26**, pp. 1056-1058, June, 1954.

HERRING, CONYERS¹

Pole of Low Energy Phonons in Thermal Conduction, *Phys. Rev.*, **95**, pp. 954-965, Aug. 15, 1954.

¹ Bell Telephone Laboratories, Inc.

HOGAN, C. L., see VAN UITERT, L. G.

KARLIN, J. E., see MUNSON, W. A.

KELLY, H. P.¹

Differential Phase and Gain Measurements in Color Television Systems, Elec. Eng., **73**, pp. 799-802, Sept., 1954.

KELLY, H. P.¹

Color Video Tester Checks Distortion, Electronics, **27**, pp. 128-131, Sept., 1954.

KENNEY, NANCY T., see FINE, M. E.

KING, R. A.,¹ and MORGAN, S. P.¹

Transmission Formulas and Charts for Laminated Coaxial Cables, Proc. I.R.E., **42**, pp. 1250-1258, Aug., 1954.

KISLIUK, PAUL¹

Arcing at Electrical Contacts on Closure — Part V. The Cathode Mechanisms of Extremely Short Arcs, J. Appl. Phys., **25**, pp. 897-900, July, 1954.

LEGG, V. E.,¹ and OWENS, C. D.¹

Magnetic Ferrites: New Materials for Modern Applications, Elec. Eng., **73**, pp. 726-729, August, 1954.

LEPOUTRE, GERARD, see DEWALD, J. F.

LOVELL, L. C., see VOGEL, F. L.

MAITA, J. P., see MORIN, F. J.

MASON, D. R.¹

Considerations on Chemical Engineering Design Problems (in French); Chimie et Industrie, **71**, pp. 477-481, March, 1954.

¹ Bell Telephone Laboratories, Inc.

McHUGH, K.³

Speculation on the Failure of Telephony, Telephony, 147, pp. 17-19, 42-43, Aug. 14, 1954.

MERRITT, F. R., see FLETCHER, R. C.

MERZ, W. J.¹

Domain Formation and Domain Wall Motions in Ferroelectric BaTiO₃ Single Crystals, Phys. Rev., **95**, pp. 690-698, Aug. 1, 1954.

MORGAN, S. P., see KING, R. A.

MORIN, F. J.,¹ and MAITA, J. P.¹

Conductivity and Hall Effect in the Intrinsic Range of Germanium, Phys. Rev., **94**, pp. 1525-1529, June 15, 1954.

MUNSON, W. A.,¹ and KARLIN, J. E.¹

The Measurement of Human Channel Transmission Characteristics, J. Acous. Soc. Am., **26**, pp. 542-553, July, 1954.

OWENS, C. D., see LEGG, V. E.

PEARSON, G. L., see FLETCHER, R. C.

READ, W. T., see VOGEL, F. L.

SCHAFER, J. P., see VAN UITERT, L. G.

SCHAWLOW, A. L.¹

Nuclear Quadrupole Resonances in Solid Bromine in Iodine Compounds, Chem. Phys., **22**, pp. 1211-1214, July, 1954.

SHOCKLEY, W., see VAN ROOSBROECK, W.

¹ Bell Telephone Laboratories, Inc.

³ New York Telephone Company.

TIEN, PING KING¹

Bifilar Helix for Backward Wave Oscillators, Proc. I.R.E., **42**, pp. 1137-1143, July, 1954.

VAN ROOSBROECK, W.,¹ and SHOCKLEY, W.¹

Photo-Radiative Recombination of Electrons and Holes in Germanium, Phys. Rev., **94**, pp. 1558-1560, June 15, 1954.

VAN UITERT, L. G.,¹ SCHAFER, J. P.,¹ and HOGAN, C. L.¹

Low Loss Ferrites for Applications at 4,000 Millicycles per Second, Letter to the Editor, J. Appl. Phys., **25**, p. 925, July, 1954.

VOGEL, F. L.,¹ READ, W. T.,¹ and LOVELL, L. C.¹

Recombination of Holes and Electrons at Lineage Boundaries in Germanium, Letter to the Editor, Phys. Rev., **94**, pp. 1791-1792, June 15, 1954.

WALKER, A. C.¹

Hydrothermal Growth of Quartz Crystals, Ind. and Eng. Chem., **48**, pp. 1670-1676, Aug., 1954.

WOLFF, P. A.¹

Theory of Secondary Electron Cascade in Metals, Phys. Rev., **95**, pp. 56-66, July 1, 1954.

WRIGHT, MARIE G.,¹ and GREMLING, R. C.¹

Xerographic Short Cut, Special Libraries, **45**, pp. 250-251, July-August, 1954.

YAGER, W. A., see FLETCHER, R. C.

¹ Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ANDERSON, P. W., MERRITT, F. R., REMEIK, J. P., and YAGER, W. A.

Magnetic Resonance in α Fe₂O₃, Monograph 2229.

BARSTOW, J. M., and CHRISTOPHER, H. N.

Measurement of Random Monochrome Video Interference, Monograph 2259.

BOND, W. L.

Notes on Solution of Problems in Odd Job Vapor Coating, Monograph 2302.

CHRISTOPHER, H. N., see BARSTOW, J. M.

CLARK, M. A.

An Acoustic Lens as a Directional Microphone, Monograph 2291.

CLOGSTON, A. M., and HEFFNER, H.

Focusing of an Electron Beam by Periodic Fields, Monograph 2267.

CORY, S. I.

A New Portable Telegraph Transmission Measuring Set, Monograph 2248.

DARROW, K. K.

Solid State Electronics, Monograph 2253.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

DITZENBERGER, J. A., see FULLER, C. S.

FINE, M. E., and KENNEY, NANCY T.

Moduli and Internal Friction of Magnetite as Affected by the Low-Temperature Transformation, Monograph 2251.

FRACASSI, R. D., and KAHL, H.

Type-ON Carrier Telephone, Monograph 2296.

FULLER, C. S., STRUTHERS, J. D., DITZENBERGER, J. A., and WOLFSTIRN, K. B.

Diffusivity and Solubility of Copper in Germanium, Monograph 2270.

GALT, J. K., YAGER, W. A., and MERRITT, F. R.

Temperature Dependence of Ferromagnetic Resonance Line Width in a Nickel Iron Ferrite: A New Loss Mechanism, Monograph 2245.

GILBERT, E. N.

Lattice Theoretic Properties of Frontal Switching Functions, Monograph 2246.

GREEN, E. I.

The Decilog: A Unit for Logarithmic Measurement, Monograph 2254.

HAGSTRUM, H. D.

Instrumentation and Experimental Procedure for Studies of Electron Ejection by Ions and Ionization by Electron Impact, Monograph 2256.

HEFFNER, H.

Analysis of the Backward-Wave Traveling-Wave Tube, Monograph 2285.

HEFFNER, H., see CLOGSTON, A. M.

JAFFE, H., see MASON, W. P.

JOHNSON, J. B., and MCKAY, K. G.

Secondary Electron Emission From Germanium, Monograph 2279.

KAHL, H., see FRACASSI, R. D.

KENNEY, NANCY, T., see FINE, M. E.

KOMPFNER, R., and WILLIAMS, N. T.

Backward-Wave Tubes, Monograph 2295.

KRETZMER, E. R.

An Amplitude Stabilized Transistor Oscillator, Monograph 2239.

LEWIS, H. W.

Search for the Hall Effect in a Superconductor I. — Experiment,
Monograph 2255.

LINVILL, J. G.

RC Active Filters, Monograph 2260.

LINVILL, J. G.

Transistor Negative-Impedance Converters, Monograph 2294.

MAITA, J. P., see MORIN, F. J.

MASON, W. P., and JAFFE, H.

Methods for Measuring Piezoelectric, Elastic, and Dielectric Coefficients of Crystals and Ceramics, Monograph 2241.

MCKAY, K. G., see JOHNSON, J. B.

MENDEL, J. T., QUATE, C. F., and YOCUM, W. H.

Electron Beam Focusing With Periodic Permanent Magnet Fields,
Monograph 2240

MERRITT, F. R., see ANDERSON, P. W.

MERRITT, F. R., see GALT, J. K.

MORIN, F. J., and MAITA, J. P.

Conductivity and Hall Effect in the Intrinsic Range of Germanium,
Monograph 2300.

MORIN, F. J., see PEARSON, G. L.

PEARSON, G. L., READ, W. T., JR., and MORIN, F. J.

Dislocations in Plastically Deformed Germanium, Monograph 2230.

PFANN, W. G.

Redistribution of Solutes by Formation and Solidification of a Molten Zone, Monograph 2290.

PIERCE, J. R.

Coupling of Modes of Propagation, Monograph 2252.

PRINCE, M. B.

Drift Mobilities in Semiconductors I, Germanium II, and Silicon,
Monograph 2271.

QUATE, C. F., see MANDEL, J. T.

READ, W. T., JR., see PEARSON, G. L.

REMEIKA, J. P.

Method for Growing Barium Titanate Single Crystals, Monograph
2247.

REMEIKA, J. P., see ANDERSON, P. W.

RYDER, R. M., and SITTNER, W. R.

Transistor Reliability Studies, Monograph 2263.

SHIVE, J. N., see SLOCUM, A.

SHOCKLEY, W., see VAN ROOSBROECK, W.

SITTNER, W. R., see RYDER, R. M.

SLOCUM, A., and SHIVE, J. N.

Shot Dependence of P-N Junction Phototransistor Noise, Monograph 2265.

SMITH, C. S.

Piezoresistance Effect in Germanium and Silicon, Monograph 2233.

SNOKE, L. R.

Observations on a Possible Method of Predicting Soil-Block Bioassay Thresholds by Distillation Characteristics of the Weathered Creosotes, Monograph 2238.

THOMAS, D. E.

A Point-Contact Transistor VHF FM Transmitter, Monograph 2234.

VALDES, L. B.

Resistivity Measurements on Germanium for Transistors, Monograph 2261.

VAN ROOSBROECK, W., and SHOCKLEY, W.

Photon-Radiative Recombination of Electrons and Holes in Germanium, Monograph 2306.

VARNEY, R. N.

Liberation of Electrons by Positive-Ion Impact on the Cathode of a Pulsed Townsend Discharge Tube, Monograph 2232.

WALKER, L. R.

Stored Energy and Power Flow in Electron Beams, Monograph 2264.

WICK, R. F.

Solution of the Field Problem of the Germanium Gyrator, Monograph 2301.

WILLIAMS, N. T., see KOMPFFNER, R.

WOLFF, P. A.

Theory of Plasma Waves in Metals, Monograph 2262.

WOLFSTRIN, K. B., see FULLER, C. S.

YAGER, W. A., see ANDERSON, P. W., and GALT, J. K.

YOCOM, W. H., see MENDEL, J. T.

Contributors to this Issue

M. M. ATALLA, B.S., Cairo University, 1945; M.S., Purdue University, 1947; Ph.D., Purdue University, 1949; Studies at Purdue undertaken as the result of a scholarship from Cairo University for four years of graduate work. Bell Telephone Laboratories, 1950-. For the past three years he has been a member of the Switching Apparatus Development Department, in which he is supervising a group doing fundamental research work on contact physics and engineering. Current projects include fundamental studies of gas discharge phenomena between contacts, their mechanisms, and their physical effects on contact behavior; also fundamental studies of contact opens and resistance. In 1950, an article by him was awarded first prize in the junior member category of the A.S.M.E. He is a member of Sigma Xi, Sigma Pi Sigma, Pi Tau Sigma, the American Physical Society, and an associate member of the A.S.M.E.

ROSEMARY E. COX, B.A., Ladycliff College, 1949; M.S. Fordham University, 1950; Bell Telephone Laboratories, 1951-. Miss Cox, who taught high school mathematics for a year before coming to the Laboratories, has been engaged in fundamental studies of contact physics. She won a New York State University Scholarship and scholarships from Ladycliff and Fordham.

GERALD V. KING, B.S., Carnegie Institute of Technology, 1920; Western Electric Company, 1921-24; Bell Telephone Laboratories, 1925-. Mr. King analyzed customer orders for step-by-step and manual systems for three years before turning to the design and checking of manual and dial PBX and community dial offices. From 1932 to 1939 he engaged in fundamental studies and development of new local and toll crossbar systems. He was involved in military work from 1939 to 1944 and since then he has been concerned with design and development of AMA accounting centers, central offices and crossbar tandem systems. He was appointed Switching Systems Development Engineer in 1952.

STEWART E. MILLER, University of Wisconsin, 1936-39; B.S. and M.S., Massachusetts Institute of Technology, 1941. Bell Telephone

Laboratories, 1941-. Since June 1954, Mr. Miller has been Assistant Director of Radio Research at Holmdel and has been in charge of research on guided systems and associated millimeter and microwave techniques. During World War II, he worked on airborne radar systems. He also worked on coaxial carrier transmissions systems. Mr. Miller holds patents in connection with automatic frequency control, an oscillator control scheme and the D-C amplifier. Member of the I.R.E., Eta Kappa Nu, Tau Beta Pi and Sigma Xi.

NORMAN A. NEWELL, E. E., Lehigh University, 1920. Mr. Newell joined Bell Telephone Laboratories in 1934 after fourteen years with the American Telephone and Telegraph Company. He has formed requirements for the development of circuit and equipment arrangements for intertoll trunk signaling systems, local and toll manual switchboards, toll switchboard trunking systems and automatic toll systems. He helped coordinate these projects and prepared descriptions of the systems for the operating companies. Mr. Newell holds patents in the fields of straightforward trunking, toll-call timing and single-frequency signaling. He is a member of the A.I.E.E., Tau Beta Pi and Pi Delta Epsilon.

WILLIAM PFERD, B.S. in M.E. Rutgers University, 1947; M.S. in M.E., Newark College of Engineering 1951. Mr. Pferd has recently been concerned with coin collector development. Previously, he worked on design and development of the station ringer for the 500-type telephone set and the dial mechanism for the same set. During World War II he served as a photographic intelligence officer in Italy with the 98th Bomb Group.

JOHN R. PIERCE, B.S., M.S., Ph.D., California Institute of Technology, 1933, 1934 and 1936; Bell Telephone Laboratories, 1936-; Appointed Director of Electronics Research, 1952. Dr. Pierce has specialized in the development of electron tubes and microwave research since joining the Laboratories. During the war he concentrated on the development of electronic devices for the armed forces. Since the war he has done research leading to the development of the beam traveling-wave tube for which he was awarded the 1947 Morris Liebmann Memorial Prize of the Institute of Radio Engineers. Dr. Pierce is the author of two books: *Theory and Design of Electron Beams*, published in second edition this year, and *Traveling Wave Tubes* (1950). He was voted the "Outstanding Young Electrical Engineer of 1942" by Eta Kappa Nu. Member of the A.I.E.E., Tau Beta Pi and Eta Kappa Nu. Fellow of the American Physical Society and the I.R.E.

ALLAN WEAVER, B.S. in E. E., University of Nebraska, 1921. Since 1945 Mr. Weaver has been in charge of a group concentrated on toll signaling with particular attention to the development of single-frequency signaling for use in connection with nationwide dialing systems. He joined Bell Telephone Laboratories in 1934 after thirteen years with American Telephone and Telegraph Company and was first concerned with the development of telegraph, telephotograph and teletypewriter systems. During World War II he was assigned to radar development. He holds thirty-seven patents in the fields of telegraphy, teletypography and signaling. Mr. Weaver is a member of the A.I.E.E., I.R.E. and Sigma Xi.



